Universidad Politécnica de Cartagena

industriales
etsii UPCT

Virtualización de un Datacenter desde el punto de vista del almacenamiento

Titulación: Ingeniería Industrial
Intensificación: Organización Industrial
Alumno: Eduardo Gómez Torre
Director: José García-Bravo García

Cartagena, 26 de Septiembre de 2016

ÍNDICE

CAPÍTULO 1.– Definición del proyecto

1.1.- Marco del proyecto

1.2.- Objetivos del proyecto

El objetivo de este Proyecto Fin de Carrera es proporcionar, de manera sencilla, un vistazo rápido a las tecnologías de almacenamiento existen e implantadas actualmente.

Primeramente se hará una breve aproximación a los sistemas de almacenamiento, definiremos sus componentes y mostraremos los diferentes tipos de topologías y redes existentes según el uso al que vayan dirigidas.

Se definirán los diferentes protocolos de acceso más utilizados hoy en día en las implantaciones de las redes SAN como son el canal de fibra y la propia red TCP/IP.

Veremos los componentes de las redes SAN desde la parte más visible, la del servidor con su sistema operativo, hasta la capa de almacenamiento y elementos de interconexión que por lo general suelen ser las capas más abstraídas de este tipo de redes para los usuarios.

Nos centraremos en las cabinas de almacenamiento y veremos de qué están compuestas. Desde la capa de acceso al servidor hasta la propia protección RAID de los grupos creados por el usuario.

Veremos los parámetros más utilizados a la hora de definir la protección de nuestros sistemas de almacenamiento y que nos ayudarán para poder decidir qué sistema, protección y tecnología implantar.

Finalmente veremos el cambio de paradigma entre el almacenamiento tradicional basado en sites activo/pasivo y el almacenamiento virtualizado con un activo/activo real que nos proporcionará alta disponibilidad y continuidad de negocio bajo cualquier tipo de contingencia.

CAPÍTULO 2.- Redes de almacenamiento

2.1.- Soportes físicos de almacenamiento

2.1.1.- Disco Duro

Es el medio de almacenamiento por excelencia desde que en 1955 saliera el primer disco duro. Los discos duros se emplean en computadores de escritorio, portátiles y unidades de almacenamiento destinado al mundo enterprise. Es el componente que se encarga de almacenar todos los datos en un sistema de información.

Mientras que la memoria RAM actúa como memoria de apoyo que almacena y pierde información según se van procesando datos, el disco duro almacena permanentemente la información introducida hasta que es eliminada.

El disco duro está compuesto por:

- Varios discos de metal magnetizado donde se guardan los datos.
- Un motor que hace girar los discos.
- Un conjunto de cabezales que leen la información guardada en los discos.
- Un electro-imán que mueve los cabezales.
- Un circuito electrónico de control que incluye la interfaz con la computadora y la memoria caché.
- Una caja hermética que protege el conjunto.

El número de discos depende de la capacidad del HDD y el de cabezales del número de discos x 2, ya que llevan un cabezal por cada cara de cada disco (4 discos = 8 caras = 8 cabezales).

Actualmente los discos utilizados pueden ser de 3’5” o de 2’5”, dependiendo del fabricante y tecnología utilizada:

Las cabinas de almacenamiento están compuestas por muchos de estos discos agrupados en arrays o DAEs (disk array enclosure), conectados internamente mediante interfaces de alta velocidad:

Los discos se agrupan de acuerdo a su tamaño, 3’5” en el DAE inferior y 2’5” en el DAE superior.

La tecnología utilizada hoy en día en los discos duros convencionales son:

- SATA: (Serial Advanced Technology Attachment
- FC:
- SAS (Serial Attached SCSI):
- NL-SAS:

2.1.2.- Disco sólido (EFD)

Una memoria de estado sólido es un dispositivo de que consta de una memoria no volátil en vez de los platos giratorios y cabezal de las unidades de disco duro convencionales. Al no tener piezas móviles, una unidad de estado sólido reduce drásticamente el tiempo de búsqueda y latencia.

Los *enterprise flash drives* (EFD) están diseñados para aplicaciones que requieren una alta tasa de operaciones por segundo, fiabilidad y eficiencia energética. En la mayoría de los casos, un *EFD* es un SSD con un conjunto de especificaciones superiores

Casi la totalidad de los fabricantes comercializan sus SSD con memorias no volátiles NAND flash para desarrollar un dispositivo no sólo veloz y con una vasta capacidad, sino robusto y a la vez lo más pequeño posible tanto para el mercado de consumo como el profesional. Al ser memorias no volátiles no requieren ningún tipo de alimentación constante ni pilas para no perder los datos almacenados, incluso en apagones repentinos, aunque cabe destacar que los SSD NAND Flash son más lentos que los que se basan en DRAM.

Una SSD se compone principalmente:

- Controladora: es un procesador electrónico que se encarga de administrar, gestionar y unir los módulos de memoria NAND con los conectores en entrada y salida. Ejecuta software a nivel de Firmware y es con toda seguridad, el factor más determinante para las velocidades del dispositivo.
- Caché: un dispositivo SSD utiliza un pequeño dispositivo de memoria DRAM similar al caché de los discos duros. El directorio de la colocación de bloques y el desgaste de nivelación de datos también se mantiene en la memoria caché mientras la unidad está operativa.
- Condensador: es necesario para mantener la integridad de los datos de la memoria caché, si la alimentación eléctrica se ha detenido inesperadamente, el tiempo suficiente para que se puedan enviar los datos retenidos hacia la memoria no volátil.

Los discos EFS o SSD tienen ventajas respecto a los discos mecánicos tradicionales:

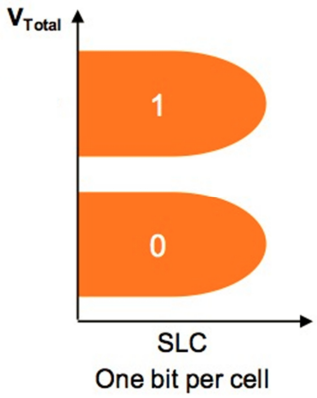
- Arranque más rápido al no tener platos que necesiten alcanzar una velocidad constante.
- Gran velocidad de escritura.
- Mayor rapidez de lectura.
- Baja latencia de lectura y escritura.
- Lanzamiento y arranque de aplicaciones en menor tiempo: resultado de la mayor velocidad de lectura y especialmente del tiempo de búsqueda.
- Menor consumo de energía y producción de calor gracias a no contar con elementos mecánicos.
- Silenciosos por la misma carencia de partes mecánicas.
- Mejora del tiempo medio entre fallos, superando dos millones de horas, muy superior al de los discos duros.
- Seguridad: permitiendo una muy rápida "limpieza" de los datos almacenados.
- Rendimiento determinista: a diferencia de los discos duros mecánicos, el rendimiento de los SSD es constante y determinista a través del almacenamiento entero. El tiempo de búsqueda permanece constante.
- El rendimiento no se deteriora mientras el medio se llena.
- Menor peso y tamaño que un disco duro tradicional de similar capacidad.
- Resistente acaídas, golpes y vibraciones sin estropearse y sin descalibrarse
- Borrado más seguro e irrecuperable de datos.

Pese a las grandes ventajas que representa el uso de discos de estado sólido también hay que contar con las siguientes desventajas:

- Coste. Los precios de las memorias flash son considerablemente más altos en relación precio/gigabyte.
- Limitada recuperación de datos ya que después de un fallo físico se pierden completamente pues la celda es destruida, mientras que en un disco duro normal que sufre daño mecánico los datos son frecuentemente recuperables.
- Fallo producido de forma inesperada: A diferencia de los discos tradicionales que empiezan a acumular sectores erróneos de forma espaciada en el tiempo, dando la posibilidad de hacer un volcado de los datos; los discos SSD producen el fallo de forma inminente sin dar tiempo a salvar ningún dato en el momento que surge el primer aviso de error.
- Vida útil: al reducirse el tamaño del transistor se disminuye directamente la vida útil de las memorias NAND. Es muy difícil de calcular su duración, ya que no depende del tiempo, sino principalmente del uso intensivo de escritura y lectura que se le dé.
- Menores tamaños de almacenamiento ofertados.
- Las tareas de mantenimiento tradicionales de los sistemas operativos acortan su vida útil de forma dramática y se recomienda su desactivación. La desfragmentación del disco duro, la utilización de memoria virtual o los procesos de indexación de búsqueda contribuyen a continuos ciclos de escritura que acortan la vida útil del SSD. Los peores procesos aplicables a una memoria de estado sólido, son los tests de rendimiento en lectura y escritura, y el formateo que desgasta automáticamente la unidad.
- Los dispositivos SSD necesitan recibir energía periódicamente, de lo contrario los datos almacenados pueden perderse. Esto hace que un corte en el suministro eléctrico, les afecte pudiendo producir la pérdida absoluta de todos los datos. Existe un método para recuperarlos que consiste en recargarlos con un ciclo completo de carga que no siempre es eficaz. Se recomienda usarlos con un dispositivo protector de la energía eléctrica SAI.

Los discos EFD pueden presentar tres diferentes tecnologías atendiendo al número celdas utilizadas en su construcción.

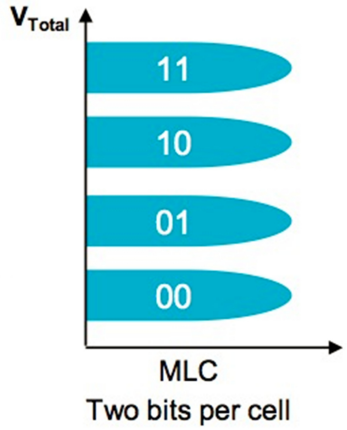
- Celdas SLC (single layer cell):



Esta tecnología consiste en cortar las obleas de silicio y obtener chips de memoria. Este proceso tiene la ventaja de que los chips son considerablemente más rápidos que los de la tecnología opuesta (MLC), mayor longevidad, menor consumo, un menor tiempo de acceso a los datos.

Como desventaja, la densidad de capacidad por chips es menor y, por tanto, un considerable mayor precio en los dispositivos fabricados con este método. A nivel técnico, pueden almacenar solamente un bit de datos por celda.

- Celdas MLC (multi-level cell):

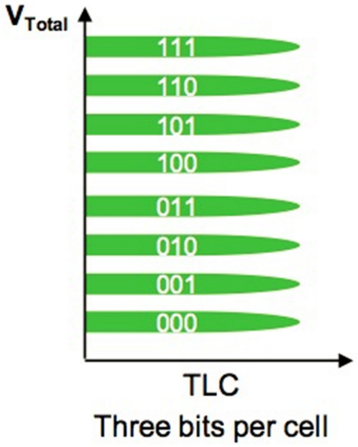


Este tecnología consiste en apilar varios moldes de la oblea para formar un sólo chip.

Las principales ventajas de este sistema de fabricación es tener una mayor capacidad por chip que con el sistema SLC y por tanto, un menor precio final en el dispositivo.

A nivel técnico es menos fiable, durable, rápido y avanzado que las SLC. Estos tipos de celdas almacenan dos bits por cada una, es decir cuatro estados, por esa razón las tasas de lectura y escritura de datos se ven mermadas.

- Celdas TLC (Triple layer cell):



Nuevo proceso en el que se mantienen tres bits por cada celda.

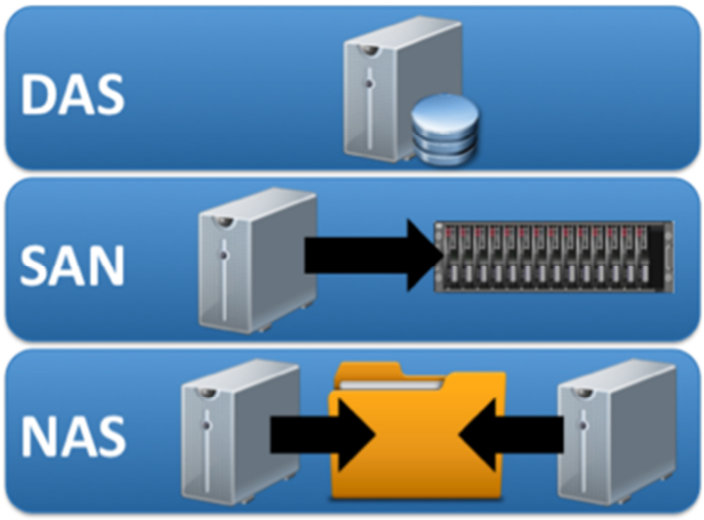
Su mayor ventaja es la considerable reducción de precio.

Su mayor desventaja es que solo permite 1000 escrituras.

2.2.- Redes de Almacenamiento

Las redes de almacenamiento son aquellas que interconectan almacenamiento externo con servidores mediante el uso de diferentes protocolos de red.

Las redes de almacenamiento se dividen en 4 tipos cuya diferencia principal es el modo de acceso al disco:



2.2.1.- DAS

El almacenamiento de conexión directa, Direct Attached Storage (DAS), es el método tradicional de almacenamiento y el más sencillo. Consiste en conectar el dispositivo de almacenamiento directamente al servidor o estación de trabajo, es decir, físicamente conectado al dispositivo que hace uso de él.

Tanto en DAS como en Storage Area Network (SAN), las aplicaciones y programas de usuarios hacen sus peticiones de datos al sistema de archivos directamente. La diferencia entre ambas tecnologías reside en la manera en la que dicho sistema de archivos obtiene los datos requeridos del almacenamiento. En un DAS, el almacenamiento es local al sistema de archivos, mientras que en un SAN, el almacenamiento es remoto.

Los protocolos principales usados en DAS son SCSI, Serial Attached SCSI (SAS) y Fibre Channel.

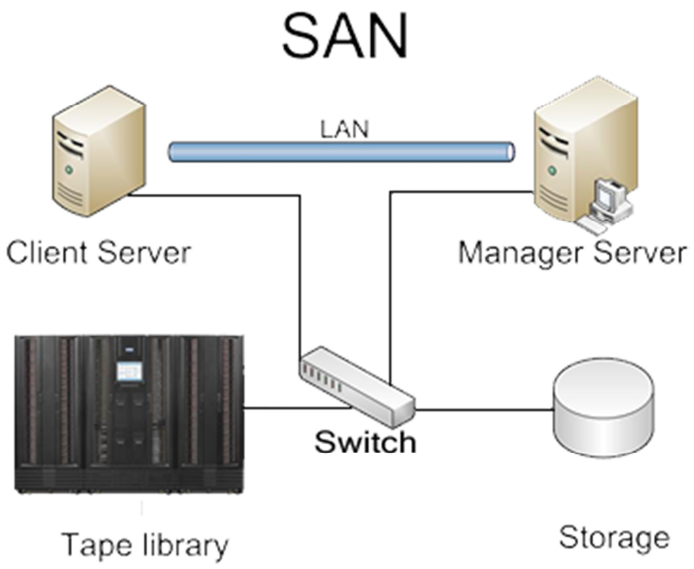
Tradicionalmente, un sistema DAS habilita capacidad extra de almacenamiento a un servidor, mientras mantiene alto ancho de banda y tasas de acceso. Un típico sistema DAS está hecho de uno o más dispositivos de almacenamiento como discos rígidos, y uno o más controladores.

Las desventajas de DAS incluyen incapacidad para compartir datos o recursos no usados con otros servidores.

2.2.2.- SAN

Las redes SAN están concebidas para proporcionar almacenamiento a nivel de bloque (LUN) a servidores, cabinas de disco y librerías de cintas.

Su función es la de conectar de manera rápida, segura y fiable los distintos elementos que la conforman y se distingue de otros modos de almacenamiento en red por el modo de acceso a bajo nivel.



Las redes SAN están basadas, principalmente, en la tecnología fibre channel (FC) aunque cada día es más habitual el uso protocolos iSCSI.

El canal de fibra permite la creación de una red SAN con conexiones de alta velocidad y numerosos dispositivos de almacenamiento.

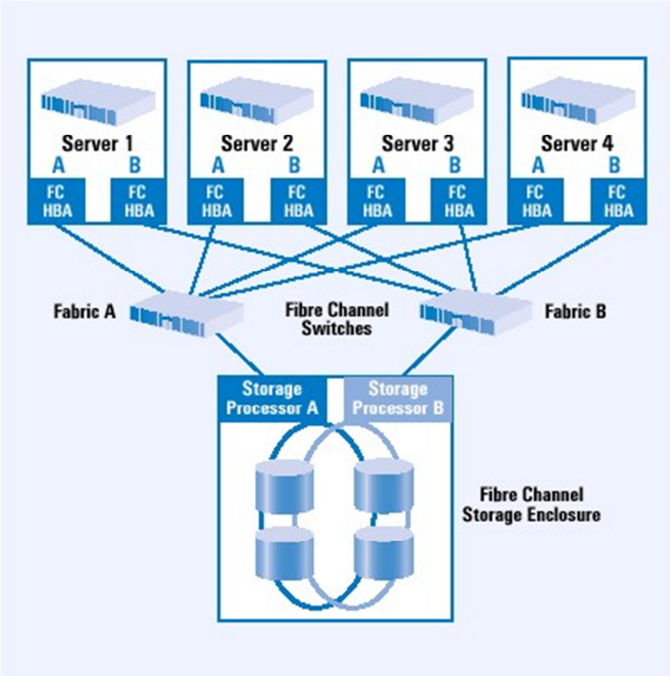
El tipo de tráfico en una SAN es muy similar al de los discos duros como ATA, SATA y SCSI. La gran mayoría de las redes SAN actuales usan el protocolo SCSI para acceder a los datos de la SAN (aunque no usen interfaces físicas SCSI) a través de FC.

Una de las principales características de las redes SAN es que son construidas para minimizar el tiempo de respuesta del medio de transmisión.

Las redes SAN permiten que múltiples servidores sean conectados al mismo grupo de discos o librerías de cintas de forma que optimizan la utilización de los sistemas de almacenamiento y los respaldos.

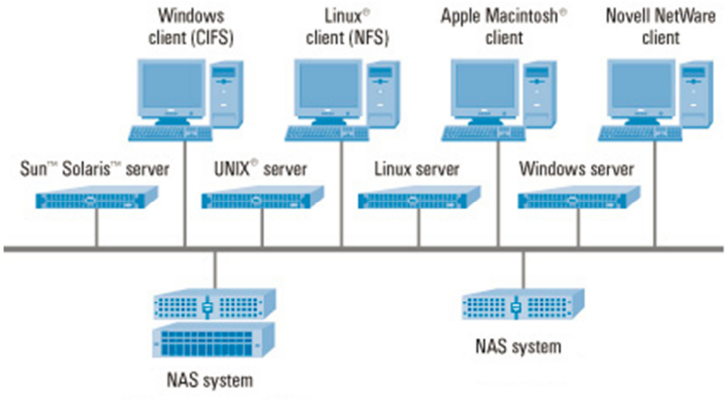
El rendimiento de cualquier sistema de computación dependerá de la velocidad de sus subsistemas. Las redes SAN han incrementado su velocidad de transferencia desde el Gigabit hasta los actualmente 2, 4, 8 y 16 Gigabits por segundo.

Una de las ventajas de las SAN es que, al tener mayor conectividad, permiten que los servidores y dispositivos de almacenamiento se conecten más de una vez a la SAN permitiendo, de esta manera, tener rutas redundantes que incrementa la tolerancia a fallos.



2.2.3.- NAS

Las redes NAS (Network Attached Storage) fueron concebidas para proporcionar almacenamiento a nivel de fichero. Difieren de las redes SAN, sobre todo, en la forma de acceso al medio que suele ser a nivel TCP/IP.



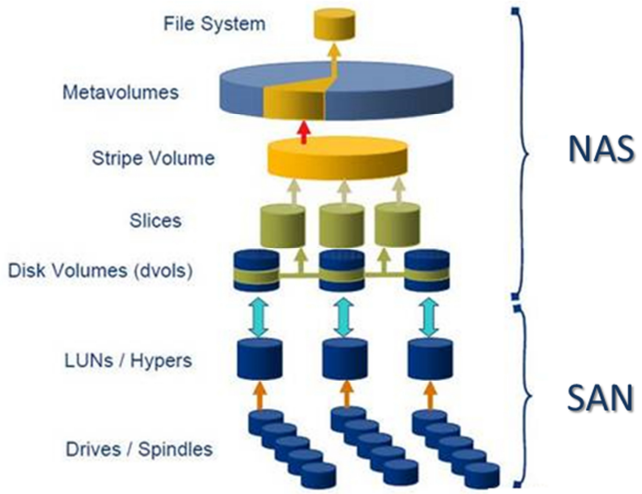
El nivel lógico más bajo usado en las redes NAS es el File System que es exportado a los diferentes sistemas que lo montan mediante los protocolos iSCSI, NFS o CIFS.

En la tecnología NAS, las aplicaciones y programas de usuario hacen las peticiones de datos a los sistemas de archivos de manera remota mediante protocolos CIFS/NFS y el almacenamiento es local al sistema de archivos.

Las ventajas principales de las redes NAS son la capacidad de compartir las unidades, un menor coste, la utilización de la misma infraestructura de red y una gestión más sencilla.

Por el contrario, las redes NAS tienen un menor rendimiento y confiabilidad por el uso compartido de las comunicaciones.

Otra de las peculiaridades de un sistema de almacenamiento NAS es que requiere de la existencia de discos para poder crear los Files Systems. Es por ello que, de forma habitual, los sistemas NAS se expandan a partir de un sistema SAN que, como vimos anteriormente, funcionan a nivel de bloque.



Los protocolos principales usados en los entornos NAS son:

- NFS (Network File System):

  Según el modelo OSI, es un protocolo de nivel de aplicación. Es utilizado para sistemas de archivos distribuido en un entorno de red de computadores de área local.

  Posibilita que distintos sistemas conectados a una misma red accedan a ficheros remotos como si se tratara de locales. Originalmente fue desarrollado en 1984 por Sun Microsystems, con el objetivo de que sea independiente de la máquina, el sistema operativo y el protocolo de transporte. Esto fue posible gracias a que está implementado sobre los protocolos XDR (presentación) y ONC RPC (sesión).

  El protocolo NFS está incluido por defecto en los Sistemas Operativos UNIX y la mayoría de distribuciones Linux.

  El sistema NFS está dividido en dos partes principales: un servidor y uno o más clientes. Los clientes acceden de forma remota a los datos que se encuentran almacenados en el servidor.

  Las estaciones de trabajo locales utilizan menos espacio de disco debido a que los datos se encuentran centralizados en un único lugar pero pueden ser accedidos y modificados por varios usuarios, de tal forma que no es necesario replicar la información.

  Los usuarios no necesitan disponer de un directorio “home” en cada una de las máquinas de la organización. Los directorios “home” pueden crearse en el servidor de NFS para posteriormente poder acceder a ellos desde cualquier máquina a través de la infraestructura de red.

  Todas las operaciones sobre ficheros son síncronas lo que significa que la operación sólo retorna cuando el servidor ha completado todo el trabajo asociado para esa operación. En caso de una solicitud de escritura, el servidor escribirá físicamente los datos en el disco y, si fuera necesario, actualizará la estructura de directorios antes de devolver una respuesta al cliente. Esto garantiza la integridad de los ficheros.

- CIFS (Common Internet File System):

  Este protocolo es utilizado principalmente en computadoras con sistemas operativos Windows y DOS. Fue inventado originalmente por IBM bajo el nombre de SMB pero la versión más común es la modificada ampliamente por Microsoft.

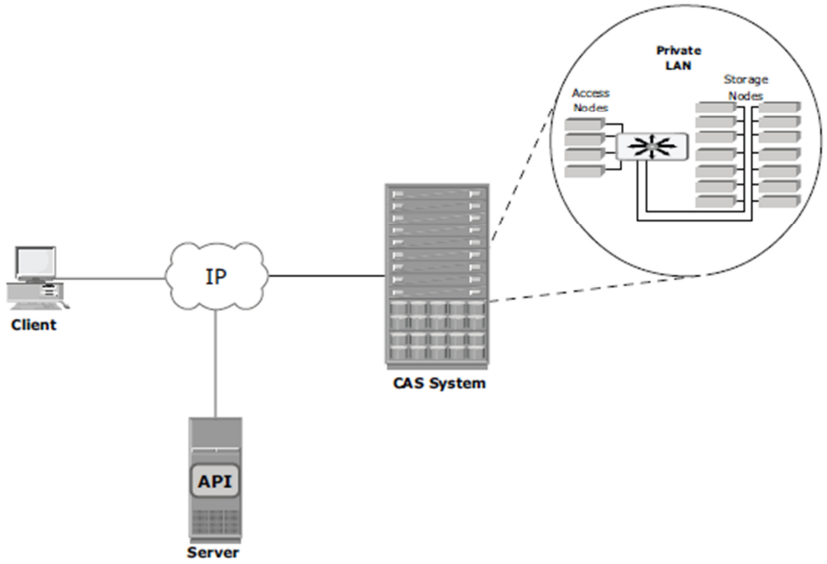
  En 1998 Microsoft renombró SMB a Common Internet File System (CIFS) y añadió más características que incluyen:

  - Soporte para enlaces simbólicos
  - Hard links
  - Mayores tamaños de archivo.

2.2.4.- CAS

Las redes CAS surgen como almacenamiento a largo plazo capaz de satisfacer los requerimientos legales respecto a integridad y conservación de los datos almacenados.

El elemento de almacenamiento en este tipo de redes es el objeto (frente al bloque o el fichero de las redes SAN y NAS)



Cuando se desea almacenar un fichero, este se transfiere por completo a la plataforma CAS de almacenamiento junto con ciertos metadatos. La plataforma genera un identificador único e irrepetible para dicho fichero en función de su contenido gracias a una función hash. Dicho identificador debe ser recordado por los programas informáticos para su posterior recuperación.

Si dicho identificador ya existe en la plataforma, significa que el documento ya está almacenado y, por tanto, no es necesario generar un duplicado. En caso contrario, el fichero es almacenado con la garantía de que jamás será sobre-escrito o modificado. Esto es debido a que cualquier otro documento, incluso pequeñas modificaciones del primero, generarán un identificador distinto. Por tanto, serían almacenados como un fichero completamente diferente.

La recuperación de los documentos tiene lugar a través del identificador único. No existen carpetas ni particiones como ocurre en los sistemas de ficheros.

El borrado o la eliminación de ficheros como tal no existe ya que no es posible hacer desaparecer un fichero una vez que ha sido almacenado. No obstante, cada fichero lleva asociado un "período de retención". Esto es, una fecha a partir de la cual el fichero no tiene valor o validez. Este es uno de los metadatos que la aplicación aporta cuando almacena el fichero. Por defecto, dicho período de retención es infinito.

Gracias a esto, la propia plataforma se encarga de eliminar los ficheros solamente cuando es pertinente. Naturalmente, se requiere la orden de "purgado" por parte de un administrador de sistemas.

Otra característica de la plataforma es que realiza periódicamente y de manera desatendida todos los chequeos necesarios para garantizar que los ficheros almacenados son legibles y permanecen íntegros, detectando y corrigiendo fallos en el medio de almacenamiento.

2.2.5.- Cloud Computing

En este tipo de computación todo lo que puede ofrecer un sistema informático se ofrece como servicio dentro del denominado IAAS (infraestructure as a service) de modo que los usuarios puedan acceder a los servicios disponibles en la nube de Internet sin conocimientos (o, al menos sin ser expertos) en la gestión de los recursos que usan.

Según el IEEE Computer Society, es un paradigma en el que la información se almacena de manera permanente en servidores de Internet y se envía a cachés

La computación en la nube son servidores desde Internet encargados de atender las peticiones en cualquier momento y se puede tener acceso a su información o servicio mediante una conexión a internet desde cualquier dispositivo móvil o fijo ubicado en cualquier lugar.

Sirven a sus usuarios desde varios proveedores de alojamiento repartidos por todo el mundo. Esta medida reduce los costos, garantiza un mejor tiempo de actividad y que los sitios web sean invulnerables a los delincuentes informáticos, a los gobiernos locales y a sus redadas policiales pertenecientes.

Cloud Computing es un nuevo modelo de prestación de servicios de negocio y tecnología que permite incluso al usuario acceder a un catálogo de servicios estandarizados y responder con ellos a las necesidades de su negocio de forma flexible y adaptativa pagando únicamente por el consumo efectuado.

El cambio que ofrece la computación desde la nube es que permite aumentar el número de servicios basados en la red. Esto genera beneficios para los proveedores que pueden ofrecer de forma más rápida y eficiente un mayor número de servicios.

Computación en nube consigue aportar estas ventajas, apoyándose sobre una infraestructura tecnológica dinámica que se caracteriza, entre otros factores, por un alto grado de automatización, una rápida movilización de los recursos, una elevada capacidad de adaptación para atender a una demanda variable, así como virtualización avanzada y un precio flexible en función del consumo realizado, evitando además el uso fraudulento del software y la piratería.

El concepto de “nube informática” es muy amplio, y abarca casi todos los posibles tipo de servicio en línea, pero cuando las empresas predican ofrecer un utilitario alojado en la Nube , por lo general se refieren a alguna de estas tres modalidades: el software como servicio (por sus siglas en inglés SaaS –Software as a Service-) , Plataforma como Servicio (PaaS) e Infraestructura como Servicio (IaaS).

2.3.- Protocolos de acceso al medio utilizados en almacenamiento

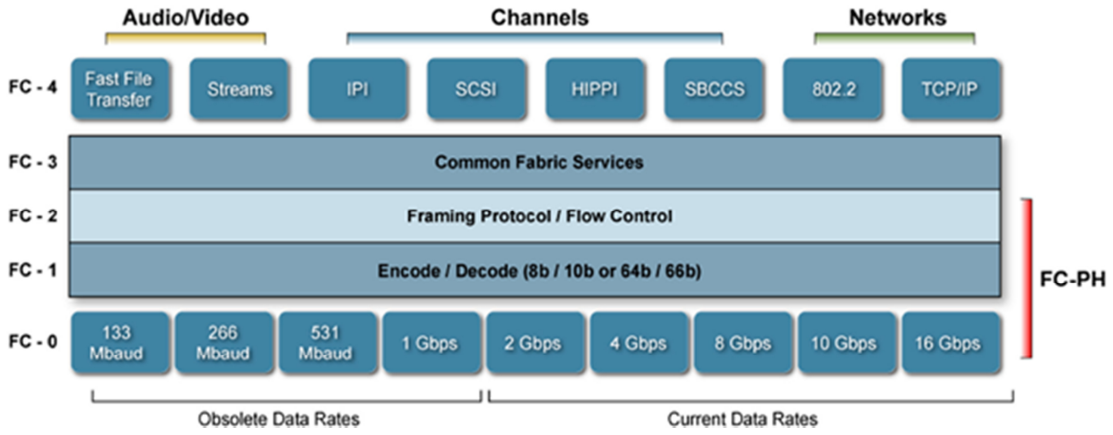
2.3.1.- Fibre Channel

El protocolo FC se caracteriza por ser un mecanismo cerrado, estructurado y predecible para la transmisión de información. Una vez que se ha establecido un canal simplemente se necesita enviar la información sin necesidad de tomar la decisión de cómo y a dónde se desea enviar. Los protocolos de canal SCSI y HPPI son los que comúnmente se utilizan para conectar dispositivos periféricos tales como unidades de disco, impresoras y unidades de cinta.

La red es impredecible y requiere de inteligencia para tomar decisiones sobre el ruteo de la información. Por lo general, esta toma de decisiones la lleva a cabo el software que hace que las redes sean mucho más lentas que los canales.

El canal de fibra combina las mejores funciones de los canales y las redes. Se trata de un estándar basado en hardware para un canal inteligente, que combina las ventajas de un canal y las tecnologías de una red en una sola interfaz de entrada/salida.

El canal de fibra aprovecha todas estas tecnologías y permite seguir utilizando protocolos establecidos funcionando a través de un cable de cobre de hasta 100 metros o de un cable de fibra óptica de hasta 10 kilómetros.



La interconexión de los nodos de una red Fibre Channel se realiza mediante tres topologías físicas:

1.- Punto a punto (Point-to-point): Conexión única entre dos nodos, todo el ancho de banda es usado por estos dos nodos.

2.- Bucle arbitrado (Arbitrated Loop): En esta topología el ancho de banda es compartida entre todos los nodos conectados al bucle. Para la conexión de todos los dispositivos de un bucle, hasta un total de 127, se utilizan concentradores (hubs).

3.- Conmutado (Switched): Permite múltiples conexiones concurrentes entre todos los nodos conectados a un conmutador o red de conmutadores (Fabric).

El estándar Fibre Channel define diferentes clases de servicio para las comunicaciones entre nodos,que se establecen para cada conexión según las necesidades de la misma a través de unos protocolos de negociación definidos entre los nodos y los elementos de la red.

Class 1: Conexión dedicada a través de la red equivalente a un enlace físico.

Class 2: Sin conexión pero con garantía de entrega de las tramas.

Class 3: Sin conexión y sin garantía de entrega. El flujo se controla en las capas superiores.

Class 4: Servicio orientado a la conexión pero con sólo un mínimo de ancho de banda garantizado.

Class 5: Servicio isócrono. No utilizado.

Class 6: Servicio multicast con conexión.

El hardware existente en el mercado implementa principalmente las clases 2 y 3.

2.3.2.- iSCSI

iSCSI (Abreviatura de Internet SCSI) es un estándar que permite el uso del protocolo SCSI sobre redes TCP/IP para la transferencia de datos.

Al contrario que otros protocolos de red diseñados para almacenamiento, como por ejemplo el canal de fibra (que es la base de la mayor parte de las redes de áreas de almacenamiento), solamente requiere una simple y sencilla interfaz Ethernet (o cualquier otra red compatible TCP/IP) para funcionar.

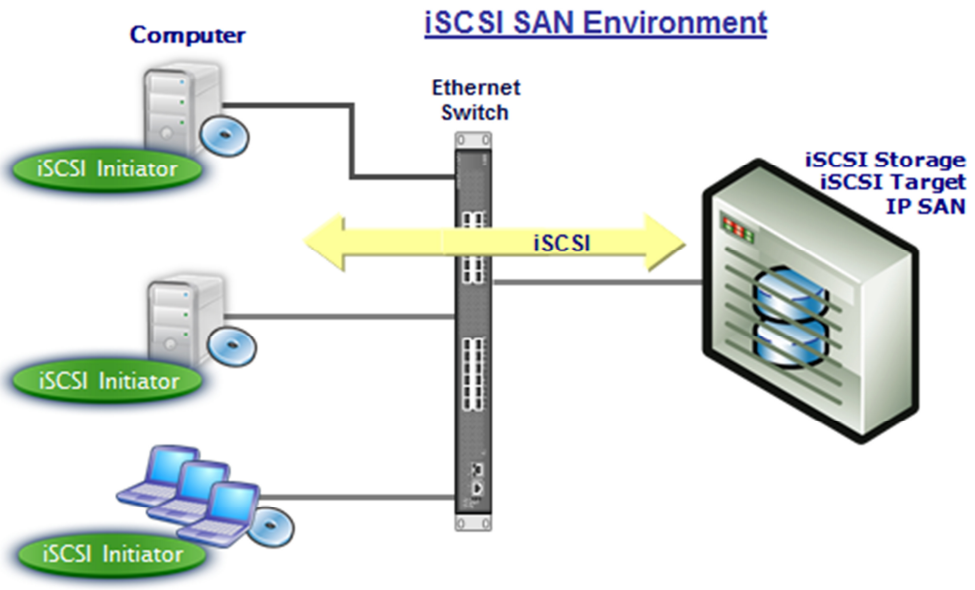
Esto permite una solución de almacenamiento centralizada de bajo coste sin la necesidad de realizar inversiones costosas ni sufrir las habituales incompatibilidades asociadas a las soluciones de canal de fibra para Redes de área de almacenamiento.

Los más críticos de iSCSI argumentan que este protocolo tiene un peor rendimiento que el canal de fibra ya que se ve afectado por la sobrecarga que generan las transmisiones TCP/IP (cabeceras de paquetes, por ejemplo).

Sin embargo las pruebas que se han realizado muestran un excelente rendimiento de las soluciones iSCSI SANs, cuando se utilizan enlaces Gigabit Ethernet.

En el contexto de almacenamiento, iSCSI permite a un host utilizar un iniciador iSCSI (initiator) para conectar a un dispositivo SCSI (target) como puede ser un disco duro o una cabina de discos en una red IP para acceder a los mismos a nivel de bloque.

Desde el punto de vista de los drivers y las aplicaciones de software, los dispositivos parecen estar conectados realmente como dispositivos SCSI locales.



2.4.- Componentes de las redes SAN

2.4.1.- Servidores

Una red de almacenamiento debe ser una red abierta y heterogénea en la que pueda entrar a formar parte todo tipo de servidores con todo tipo de sistemas operativos que puedan acceder al almacenamiento de la red.

La red entre los servidores y el almacenamiento será transparente a las aplicaciones, que verán los discos y cintas magnéticas compartidas como si fuesen dispositivos locales del sistema. Los servidores se conectan a la SAN mediante uno o varios adaptadores Fibre Channel (HBA, Host Bus Adapter).

El software del sistema operativo deberá estar optimizado para la nueva situación: número elevado de dispositivos y varios caminos distintos alternativos para acceder al mismo dispositivo con reconfiguración automática en caso de caída de un camino, sistemas de ficheros adaptados, capacidad de arranque desde un disco de la SAN, etc.

Los servidores corren diferentes sistemas operativos, que no es más que un programa o conjunto de programas de un sistema informático que gestiona los recursos de hardware y provee servicios a los programas de aplicación de software.

Los sistemas operativos se clasifican, de forma general, en:

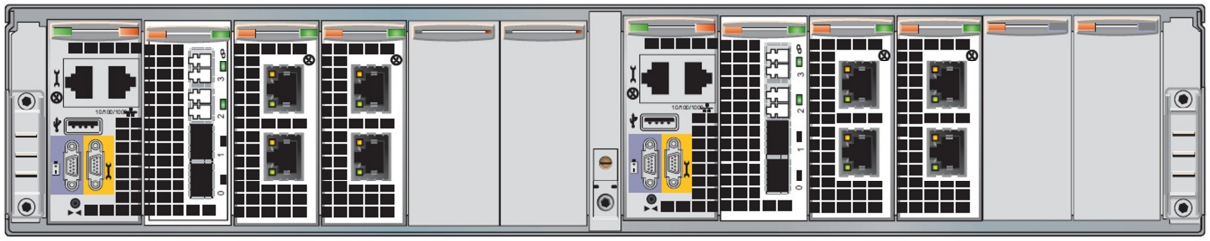
- Sistemas UNIX
- Sistemas Windows
- Sistemas virtualizados (VMWARE, HyperV)

2.4.2.- Cabinas de almacenamiento

Los dispositivos de almacenamiento son la base de la SAN. La SAN permite liberar el almacenamiento de tal manera que ya no forma parte de un bus particular de un servidor si no que se distribuye a través de caminos o paths e incluso a través de sites separados entre sí.

Las cabinas de discos se diseñan teniendo en cuenta la importancia de la disponibilidad y seguridad de los datos contenidos en sus dispositivos mediante elementos redundados e intercambiables en caliente: controladoras, módulos de caché, baterías, fuentes de alimentación, discos.

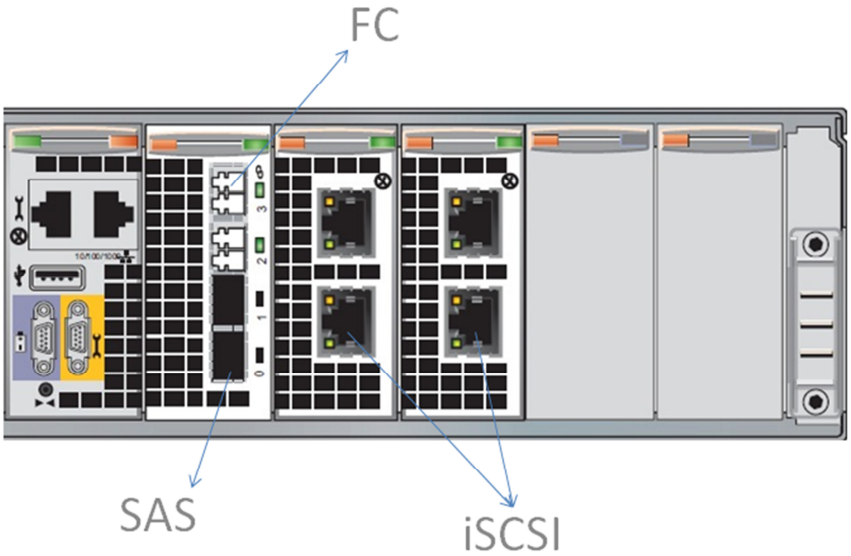
El componente fundamental de una cabina de discos es su controladora, redundada con su pareja:



Las controladoras se conectan a la SAN mediante puertos Fibre Channel y a la estructura interna de la cabina mediante buses SCSI o conexiones Fibre Channel internas formándose un doble bucle balanceado.

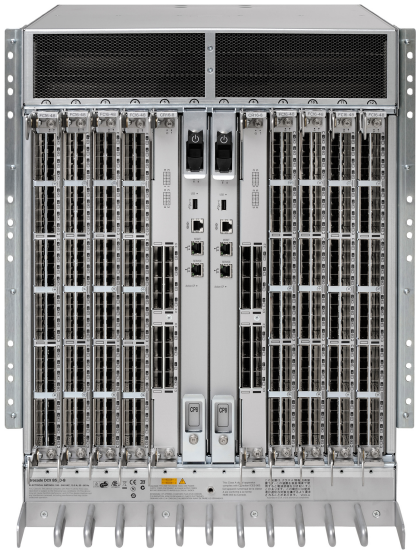
Los discos serán dispositivos Fibre Channel. Las controladoras tienen funcionalidades de redundancia y paridad tipo RAID de acceso a los volúmenes o LUN por ellas gestionadas (LUN Masking) y con capacidad de tomar el control del sistema transparentemente si su pareja falla.

Las controladoras de almacenamiento están formadas por diferentes módulos que proporcionan la comunicación al medio físico que más se adapte a las necesidades de la infraestructura.

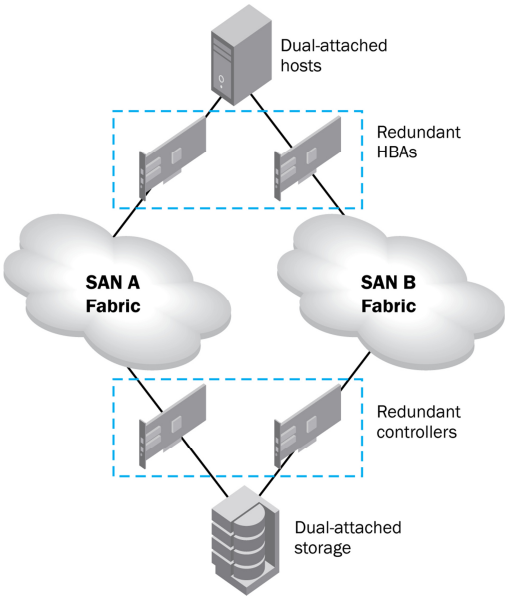


2.4.3.- Elementos de interconexión – Switches

Un switch o conmutador es un equipo de red de alto rendimiento capaz de interconectar muchos dispositivos o interactuar con otros conmutadores.



Al conjunto de conmutadores de una red se denomina fabric o switch fabric. Cualquier dispositivo conectado a un puerto de un switch puede conectarse con cualquier otro dispositivo de la red y será la infraestructura de conmutadores la encargada de encaminar todo el tráfico de un dispositivo a otro.



Como método de control de acceso entre dispositivos dentro de un Fabric se usa el port zoning que permite realizar mapeos lógicos entre puertos a los cuales se les permite tener visibilidad y tráfico.

2.4.3.1.- GBICs

Una interfaz Gigabit Converter (GBIC, por sus siglas en inglés) es un módulo cuya función es aumentar la velocidad de transferencia de datos a través de una red.

Actúa como transmisor-receptor convirtiendo las corrientes eléctricas en señales ópticas, antes de cambiar las señales ópticas en corrientes eléctricas digitales.

El GBIC se creó para simplificar la conexión y el diseño del Switch. Cada módulo GBIC está en su lugar para hacer más fácil la administración del sistema de redes de comunicación electro-ópticas.

2.4.4.- Aplicaciones de las redes de almacenamiento

Las aplicaciones de una red de almacenamiento proporcionan mejoras en el rendimiento, en la gestión y en la escalabilidad de las infraestructuras de las tecnologías de la información. El hecho de que servidores y sistemas de almacenamiento compartan la misma red permite la transferencia de datos de tres maneras distintas:

a) Entre servidor y almacenamiento. Es el modelo tradicional de interacción, aunque en el caso de una SAN el mismo dispositivo de almacenamiento puede ser accedido por múltiples servidores.

b) Entre servidores. La propia SAN puede usarse como medio de comunicaciones entre servidores.

c) Entre dispositivos de almacenamiento.

La SAN permite la transferencia de datos entre sistemas de almacenamiento sin intervención directa de los servidores. Son diversas las aplicaciones de una SAN. A continuación se enumeran algunas de ellas.

2.4.5.- Gestión centralizada

Una SAN permite agrupar los dispositivos de almacenamiento formando elementos especializados y separados de los servidores. Ya no necesitamos una tarjeta RAID y varios discos para cada servidor, con una cabina de discos y varios servidores en una SAN optimizamos la gestión del almacenamiento.

La interconexión de todo el almacenamiento dentro de la misma infraestructura de red permite la utilización de las técnicas de gestión globales propias de las redes en una SAN.

2.4.6.- Compartición de datos

En el caso anterior obtenemos el beneficio de una mejora en la utilización de los dispositivos de almacenamiento, pero seguimos teniendo un modelo en el que cada volumen de almacenamiento es asignado a un único servidor.

Tanto el protocolo SCSI como el software de sistema operativo está adaptado para estos casos, todavía no contemplan la situación en que un conjunto de servidores compartan simultáneamente un mismo volumen. Para obtener una verdadera compartición de datos en necesario introducir inteligencia en los elementos de la SAN.

Las soluciones propuestas van desde el nivel hardware (ampliaciones del estándar SCSI en la parte del almacenamiento para la gestión de los bloqueos), al nivel del sistema operativo (sistemas de ficheros distribuidos de nueva generación, servidores de bloqueos distribuidos, software de clúster) hasta el nivel de aplicación (gestión del almacenamiento por la propia aplicación de base de datos)

2.4.7.- Protección de datos

La oportunidad de conectar unidades de cintas magnéticas a una SAN ofrece la posibilidad de reducir la carga de la red de datos con el tráfico de copias de seguridad. Incluso si los elementos de la red tienen la funcionalidad adecuada se pueden realizar las copias de disco a cinta sin pasar por el servidor (server-free backup).

Con la existencia de una SAN se pueden implementar nuevos sistemas de protección: mirroring entre dos volúmenes remotos, snapshots de volúmenes como paso intermedio para un volcado a cinta, etc.

2.4.8.- Alta disponibilidad

Una SAN permite que varios servidores tengan acceso al mismo volumen de datos por uno o varios caminos dependiendo de la topología y configuración de la misma. Es el escenario adecuado para los entornos críticos y para la implementación de clústers de servidores.

2.4.9.- Continuidad de negocios

Una SAN puede extenderse a largas distancias e incluso su tráfico puede ser encaminado a través de otras redes de área extensa. Esto permite soluciones de recuperación ante desastres.

2.4.10.- Ventajas y desventajas

El rendimiento de la SAN está directamente relacionado con el tipo de red que se utiliza. En el caso de una red de canal de fibra, el ancho de banda es de aproximadamente 100 megabytes/segundo (1.000 megabits/segundo) y se puede extender aumentando la cantidad de conexiones de acceso.

La capacidad de una SAN se puede extender de manera casi ilimitada y puede alcanzar cientos y hasta miles de terabytes.

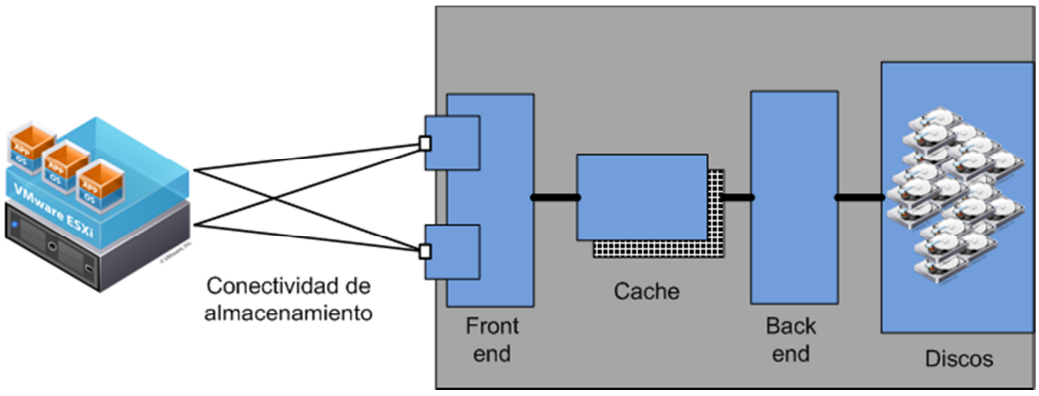
Una SAN permite compartir datos entre varios equipos de la red sin afectar el rendimiento porque el tráfico de SAN está totalmente separado del tráfico de usuario. Son los servidores de aplicaciones que funcionan como una interfaz entre la red de datos (generalmente un canal de fibra) y la red de usuario (por lo general Ethernet).

Por otra parte, una SAN es mucho más costosa que una NAS ya que la primera es una arquitectura completa que utiliza una tecnología que todavía es muy cara. Normalmente, cuando una compañía estima el TCO (Coste total de propiedad) con respecto al coste por byte, el coste se puede justificar con más facilidad.

CAPÍTULO 3.- El almacenamiento desde el punto de vista conceptual

3.1. – Componentes de una red SAN desde el almacenamiento

Se tienen 4 componentes principales en un sistema de almacenamiento: frontend, cache, backend y discos.



El rol de cada uno de los componentes es el siguiente:

3.1.1.- Front End

Es el encargado de manejar la interacción directa con el host que está pidiendo acceso de cualquier tipo a los datos almacenados, ya sea lectura o escritura. Generalmente consiste de dos o más controladoras que, a su vez, tienen múltiples puertos para poder permitir la conexión de muchos servidores a la vez.

La controladora es la encargada de manejar el protocolo utilizado en la comunicación (iSCSI, FC, FCoE, NFS…). Esta controladora se comunica con la caché mediante un bus de información interno a través del cual envía escrituras y lecturas.

Una vez que la caché da acceso a la información (ya sea entregándola a partir de su memoria o acceso directo al backend) se envía un ACK (acknowledge) para confirmar que la operación síncrona ha sido satisfactoria.

3.1.2.- Cache

La cache generalmente está compuesta por memoria DRAM, aunque en los últimos sistemas de almacenamientos más avanzados pueden llegar a combinar distintos niveles de cache añadiendo a la memoria DRAM la utilización de discos SSD.

Este componente es crítico para poder acelerar las operaciones de lectura y escritura almacenando aquella información que ha mostrado una actividad (I/O) mucho más alta en la memoria propia de este cache para que los servidores que tienen acceso al sistema de almacenamiento reciban dicha información a partir de este cache y prevenir el acceso a los discos duros mecánicos que tienen un tiempo de acceso mucho más elevado.

La cache está constituida por dos componentes principales data store y tag ram. La función del data store es básicamente almacenar la información en unidades llamadas páginas mientras que la función de tag ram es la darle seguimiento para saber dónde se encuentra la información y así poder utilizarla.

Es importante saber que en el momento que un servidor pide información al sistema de almacenamiento se busca dicha información en el tag ram para determinar si será servida o no desde cache.

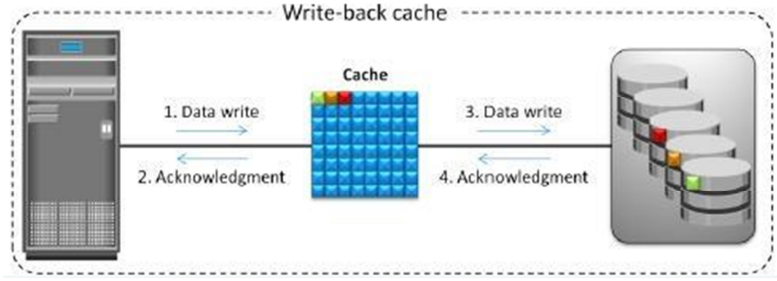
En el caso de que esta sea servida desde cache se le conoce como un Read Hit (acierto de caché), si la información no se encuentra en cache esta tendrá que ser servida a partir de los discos a través del backend y esto es conocido como Read Miss (no acierto de caché).

En el momento que se detecta un acceso secuencial puede darse que el sistema de almacenamiento cuente con la tecnología de prefetch (este puede ser de tamaño fijo o variable) o read ahead a partir de la cual se definirá un conjunto de bloques o información secuencial a la información que se está accediendo para ser transferido de los discos hacia el cache para que en el caso que el host los requiera se pueda acelerar el acceso a esta información.

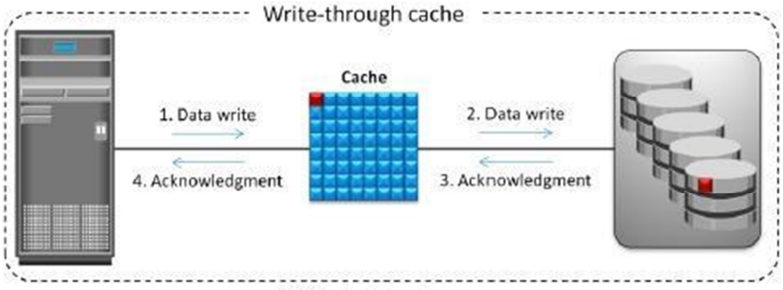
A mayor Read Hits mejor rendimiento se tendrá en las operaciones de lectura hacia nuestro sistema de almacenamiento.

La cache no solo optimiza las operaciones de lectura ya que se tienen dos distintos modos de cache para las operaciones de escritura:

1. Write-back cache: Permite tener tiempos de respuesta más rápidos debido a que en el momento que llega una operación de escritura esta es almacenada en la cache e inmediatamente se envía un acknowledge al host. Posteriormente, esta escritura se enviará a disco pero se hará en un momento en el que no sea crítico el tiempo de respuesta hacia el host.

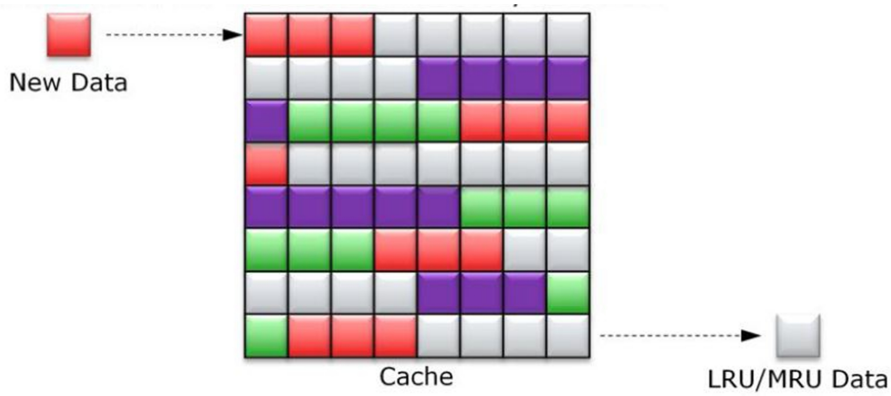


2. Write-through: Cuando se recibe la escritura se escribe inmediatamente a disco. Tras la escritura, se envía el acknowledge al host. Al tratarse de una escritura directa en disco, el tiempo de respuesta es bastante más elevado que en el modo anterior.



Las operaciones que son almacenadas en ambos modos de cache viene determinada por el tamaño máximo de I/O configurado mediante la cual se decide si esta IO se almacenará en cache o será enviada directamente a disco.

La cantidad de cache en cualquier sistema de almacenamiento es mucho menor que el espacio disponible en disco por lo que debe ser utilizada de la manera más óptima posible para almacenar solamente la información más crítica para así poder optimizar las operaciones de I/O.

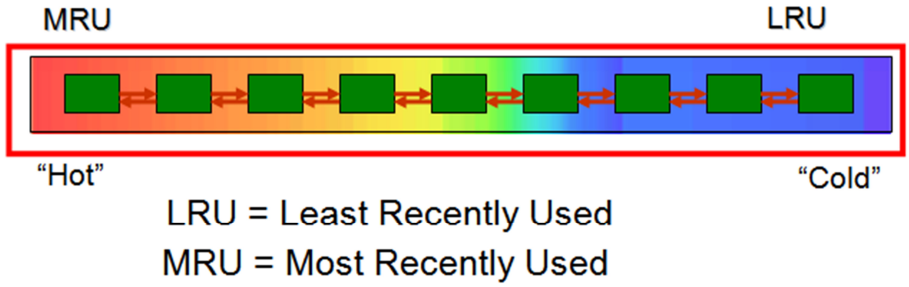


Para evitar la saturación de la caché y poder gestionar de la mejor manera posible los datos que en ella se mantienen se hace uso de los siguientes algoritmos:

1. Least Recently Used (LRU): Este algoritmo se encuentra constantemente monitoreando el último acceso a la información identificando las páginas almacenadas en cache que no han sido utilizadas durante un largo intervalo de tiempo.

   En el momento en que dichas páginas son identificadas se marcan para ser reutilizadas o, en el caso de que estas contengan información de escritura que no ha sido enviada a disco, enviarlas para después marcar dichas páginas como reutilizables.

2. Most Recently Used (MRU): Este algoritmo es exactamente lo opuesto a LRU. Se definen las páginas de cache que han sido utilizadas recientemente tomando como premisa que, debido a que la información fue accedida hace poco tiempo, esta no será requerida pronto.

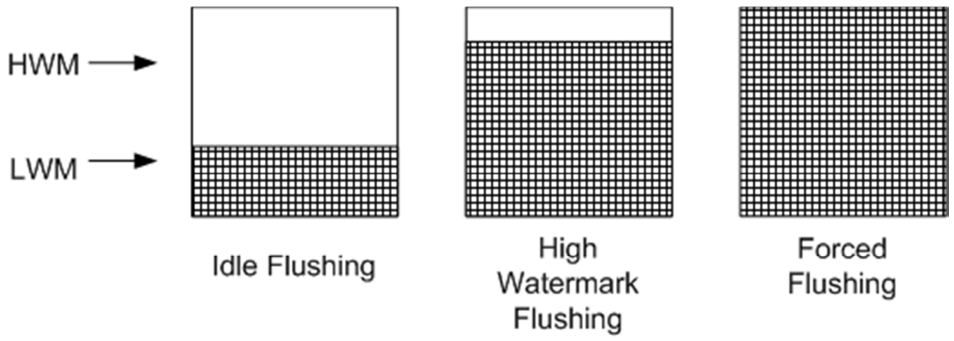


Independientemente del algoritmo para mantener almacenada o no la información en cache se requiere definir en qué puntos de utilización (espacio de cache consumido) se deberá realizar un flush de toda aquella información que se encuentra en cache y no ha sido escrita a disco para poder gestionar el espacio disponible en cache y no saturarlo.

Basándose en la cantidad de espacio consumido de la cache se definirá el ritmo de flush o escritura de información a partir del cual los datos cacheados se bajarán a disco.

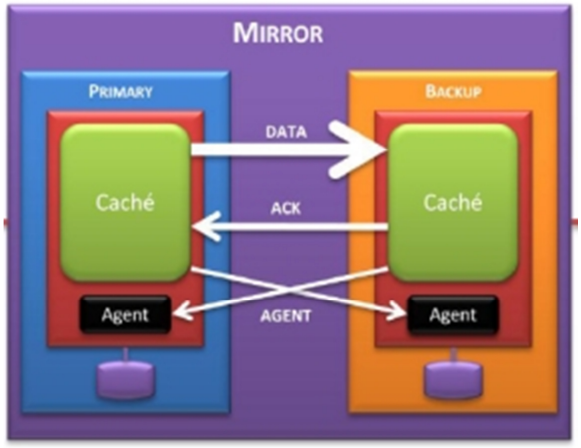
Se tienen 3 estados:

1. High Watermark: El flush ocurre a un ritmo más rápido debido a que ya se superó la cantidad de páginas de cache marcadas como límite por el usuario.
2. Idle flushing: El flush sucede a un ritmo moderado debido a que está por debajo del nivel de utilización marcado como Low Watermark
3. Forced Flushing: Estamos antes el caso más crítico de uso de la caché, ya que se ha llegado a un 100% de consumo de espacio por lo que se requiere un ritmo de escritura de la información en cache a disco muy rápida.



Hemos visto la importancia de la caché en los sistemas de almacenamiento y surge por ello la necesidad de protegerla antes posibles fallos. Para ello se cuenta con la ayuda de dos técnicas:

- Caché Mirroring: Como se ha descrito anteriormente, los sistemas de almacenamiento cuentan con controladoras duplicadas. Cada una de ellas dispone de su propia caché que se encuentra en mirror. Es decir, toda aquella escritura que no se ha bajado todavía a disco reside en las dos controladoras de manera que si una entra en fallo el dato no se pierde.



- Cache Vaulting: Este método logra proteger la información que vive en el cache a través de discos dedicados a almacenar la información que este activa en el cache, por lo que cuando se tiene un problema de energía eléctrica la información es copiada a estos discos dedicados y definidos como vault donde la información deja de ser volátil. En el momento en el que la energía eléctrica regresa esta información es copiada de nuevo a la cache y es escrita a los discos.

3.1.3.- Back End

Este componente del sistema de almacenamiento sirve como interfaz de comunicación entre la cache y los discos físicos. El flujo de información sucede de la siguiente manera:

1. La operación de escritura es enviada a disco desde la cache
2. Llega al Backend donde puede ser almacenada durante un breve periodo de tiempo antes de ser enviada a los discos.

En el Backend también se cuenta con mecanismos para detectar errores y corregirlos además de los mecanismos para controlar el o los niveles de RAID.

En el Backend todos los elementos están redundados para así poder prevenir de puntos únicos de fallo.

3.1.4.- Discos Físicos

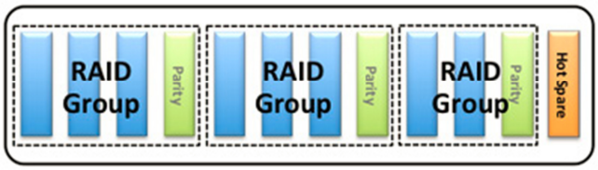
Estos son los discos tal y como los conocemos. Aquí es donde sucede todo la asignación y reserva de espacio según distintas políticas definidas donde podemos estar hablando de alta disponibilidad, rendimiento, capacidad…

Para asignar espacio primero debemos segmentar o agrupar los discos físicos en grupos lógicos a los cuales se les asigna un nivel de RAID y que llamaremos Raid set o Raid Group.

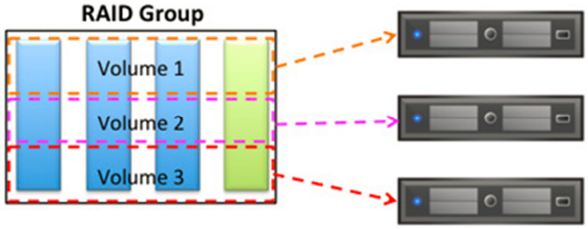
Una vez creado el RAID Set este es particionado con lo que se crean las unidades lógicas que estarán siendo entregadas a los servidores que requieren acceso a disco, estas unidades se les conoce como LUNs o Logical Unit Number, a cada unidad se le asigna un número único.

3.2.- Raid Groups

Un Raid Group (RG) es un conjunto de discos que comparten una determinada protección RAID.



Dentro de los Raid Groups se crean las LUNs que son configuradas de manera stripeada a partir de todos los discos que componen el grupo, de manera que la protección asignada es la que le corresponde a cada una de las LUNs.



De forma general, en los Raid Groups tradicionales no se pueden mezclar tecnologías de discos. Tampoco son ampliables, por lo que una vez que son creados a partir de una serie de discos deben perdurar en el tiempo.

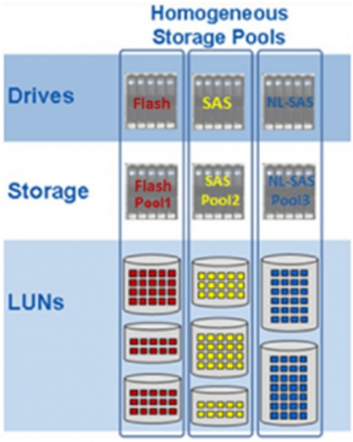
3.3.- Storage Pools

Los Storage Pools siguen el concepto de los Raid Groups, esto es, un conjunto de discos que comparten una protección específica y a partir del cual se van creando los "cortes" o LUNs.

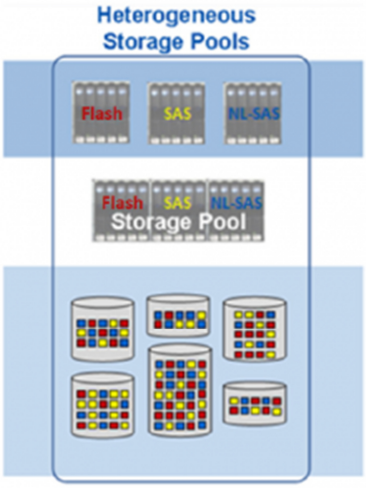
Internamente estos pools están compuestos por Raid Groups privados. La principal ventaja que presentan los Storage pools es que pueden ampliarse en caliente simplemente añadiendo nuevos raid groups privados.

Una de las mayores ventajas sobre los Raid Groups tradicionales es la de poder mezclar diferentes tecnologías de disco y poder jugar, de esa manera, con los denominados Tiers. De esta forma encontramos dos tipos de pools:

1. Pools homogéneos. Aquellos formados por el mismo tipo de discos a nivel interno, recomendados para aquellas aplicaciones con unos requerimientos de performance conocidos y similares.



2. Pools heterogéneos: Aquellos formados por diferentes tipos de discos a nivel interno y que nos permitirá la utilización de tecnologías de tier.

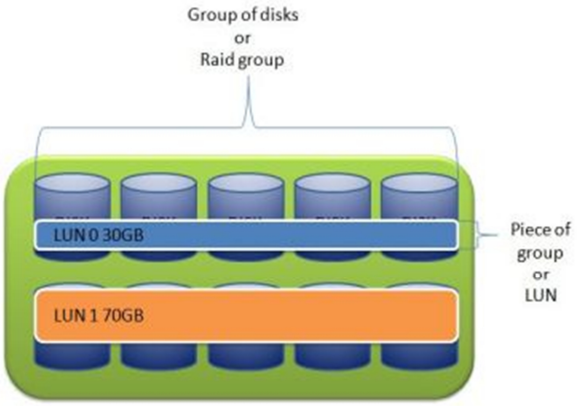


La tecnología de auto-tier permite a la cabina de almacenamiento analizar el uso de los discos pudiendo sacar estadísticas de aquellos chunks más utilizados de forma que sea capaz de mover los bloques más utilizados a discos rápidos y bajar los que menos uso tienen a disco más lento, mejorando el performance de las aplicaciones.

Cada fabricante establece una granularidad en lo que se refiere al tamaño de los bloques que se mueven que puede ir desde los 64GB a los 256GB.

3.4.- LUNs

Una LUN (logical unit number) es un particionado lógico de los discos que conforman un Raid Group, Raid Set o Storage Pool. Para la creación de una LUN se coge un pequeño trozo de cada disco de forma que la protección configurada para el grupo completo le pueda aplicar a esta entidad lógica.



Las LUNs se exportan a los servidores o hosts a través de un proceso denominado masking mediante los protocolos iSCSI o Fibre Channel.

A nivel de host la LUN es un disco conectado a él como si se tratara de un disco local cualquiera.

Cuando se crea una LUN se le asigna un identificador a nivel cabina, único para ella, conocido como LUN ID o ALU. Este número puede encontrase en decimal o hexadecimal, siempre dependiendo del fabricante. Además, a cada una de la LUN, se le asigna una WWN (world wide name) que es única para esa LUN.

Esta WWN es un número identificado formado por 8 ó 16 bytes que está controlado por una entidad física denominada NAA (Network Address Authority).

Cuando se crea una LUN obtenemos dicho identificador y, al presentarla al host, este verá el mismo id. Un ejemplo del listado de discos desde un host donde se pueden ver los identificadores:

```
ls -l /dev/disk/by-id/
[…]
lrwxrwxrwx 1 root root  9 Jul  4 22:00 wwn-0x5002e10000000000 -> ../../sr0
lrwxrwxrwx 1 root root  9 Jul  4 22:00 wwn-0x500277a4100c4e21 -> ../../sda
lrwxrwxrwx 1 root root 10 Jul  4 22:00 wwn-0x500277a4100c4e21-part1 -> ../../sda1
lrwxrwxrwx 1 root root 10 Jul  4 22:00 wwn-0x500277a4100c4e21-part2 -> ../../sda2
lrwxrwxrwx 1 root root 10 Jul  4 22:00 wwn-0x500277a4100c4e21-part3 -> ../../sda3
```

A cada fabricante se le asignan 3 bytes para que los puedan asignar a sus sistemas de almacenamiento a la hora de crear las LUNs de manera que puede resultar relativamente sencillo identificar a qué cabina de almacenamiento pertenece una LUN dada:

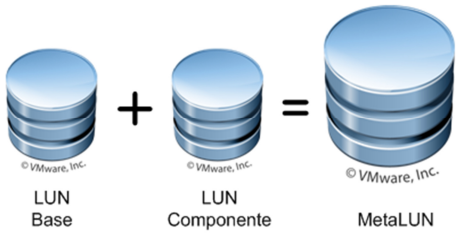
- 00:10:86 ATTO Technology
- 00:60:69 Brocade Communications Systems
- 00:05:1E Brocade Communications Systems, acquired with Rhapsody Networks
- 00:60:DF Brocade Communications Systems, acquired with CNT Technologies Corporation
- 08:00:88 Brocade Communications Systems, acquired with McDATA Corporation. WWIDs begin with 1000.080
- 00:05:30 Cisco
- 00:05:73 Cisco
- 00:05:9b Cisco
- 00:D3:10 Dell, Inc., for Dell Compellent Storage products
- 00:01:E8 Dell, Inc., for Dell Force10 Networking Products
- 00:23:29 DDRdrive LLC, for DDRdrive X1
- 00:60:16 EMC Corporation, for CLARiiON/VNX
- 00:60:48 EMC Corporation, for Symmetrix DMX
- 00:00:97 EMC Corporation, for Symmetrix VMAX
- 00:01:44 EMC Corporation, for VPLEX
- 00:00:C9 Emulex
- 00:60:B0 Hewlett-Packard - Integrity and HP9000 servers. WWIDs begin with 5006.0b0
- 00:11:0A Hewlett-Packard - ProLiant servers. Formerly Compaq. WWIDs begin with 5001.10a
- 00:01:FE Hewlett-Packard - EVA disk arrays. Formerly Digital Equipment Corporation. WWIDs begin with 5000.1fe1 or 6000.1fe1
- 00:17:A4 Hewlett-Packard - MSL tape libraries. Formerly Global Data Services. WWIDs begin with 200x.0017.a4
- 00:0C:CA HGST, a Western Digital Company
- 00.60:E8 Hitachi Data Systems
- 00:50:76 IBM
- 00:17:38 IBM, formerly XIV.

3.5.- MetaLUNs

Antes de la existencia de los Storage Pools no era posible aumentar el tamaño de las LUNs que se había creado dentro de un Raid Group o Raid set.

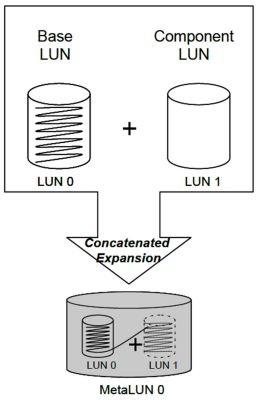
Por entonces se utilizaba el concepto de MetaLUN que permitía expandir una LUN existente ya sea para darle mejor rendimiento o ampliar su espacio.

Una MetaLUN básicamente es la construcción lógica a partir de dos o más LUNs. Para ello es necesario contar con la LUN base (LUN que generalmente tiene el dato) y ampliarla con una LUN componente:

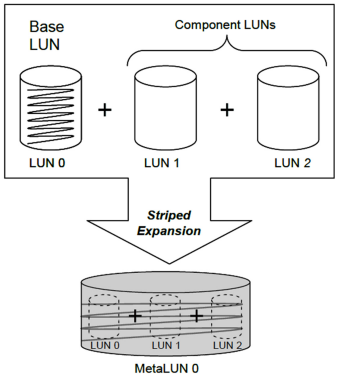


A la hora de crear una MetaLUN se pueden utilizar dos técnicas diferentes:

1. Concatenación: La LUN componente se colocaba inmediatamente después de la LUN base. La creación de esta nueva LUN era instantánea y únicamente se conseguía ampliar el espacio total.



2. Stripped: La LUN base y la LUN componente se stripean entre sí para crear una nueva entidad lógica con la suma de las dos capacidades pero con los datos mezclados entre sí. Esta forma de creación permitía, además del aumento de capacidad, la ganancia en rendimiento al contar esa nueva LUN con más cabezales para la lectura/escritura de los datos.



3.6.- LUN Masking

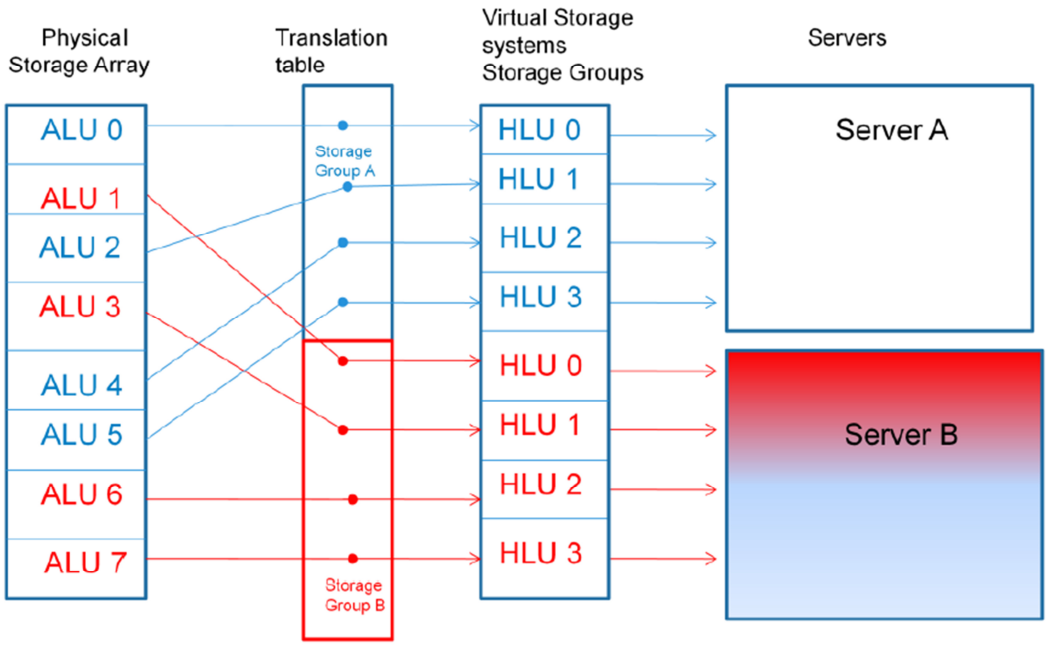
Para poder presentar las LUNs disponibles a los servidores se deben definir que hosts tienen acceso a los LUNs. A este proceso se le conoce como LUN Masking.

En el momento de realizar LUN masking debemos definir las LUNs y los identificadores de los Hosts según el protocolo que estemos utilizando.

Cada vez que se asigna una LUN a un host se requiere la configuración de un nuevo parámetro llamado Host ID. Este identificador no es más que una referencia a la LUN, en decimal o hexadecimal, que será usado por el host como la posición scsi dentro de su estructura. Dentro de un mismo Host Group o Storage Group no es posible repetir el mismo host ID para dos LUNs diferentes.

A modo de resumen, en la figura siguiente, tenemos en color rojo un Storage Group A y en azul un Storage Group B

1. Physical Storage Array: Hace referencia a la ALU o LUN id
2. Storage Group: Hace referencia al host LUN.



3.7.- Protecciones de discos en los sistemas de almacenamiento

Los discos utilizados en los sistemas de almacenamiento se agrupan, como hemos visto anteriormente, en pools y Raig Groups. Para cada una de las agrupaciones se puede configurar una protección denominada RAID que permiten que un disco falle mecánicamente y que aun así los datos del conjunto sigan siendo accesibles para los usuarios.

En lugar de exigir que se realice una restauración costosa en tiempo desde una cinta, DVD o algún otro medio de respaldo lento, un RAID permite que los datos se recuperen en un disco de reemplazo a partir de los restantes discos del conjunto mientras, al mismo tiempo, permanece disponible para los usuarios en un modo degradado.

El tiempo de no disponibilidad suele tener graves repercusiones económicas para el negocio de las empresas y por ello la utilización de configuraciones RAID se ha convertido en una necesidad para ellas.

El uso de RAID también puede mejorar el rendimiento de ciertas aplicaciones.

Los niveles RAID 0, 5 y 6 usan stripping de datos lo que permite que varios discos atiendan simultáneamente las operaciones de lectura aumentando la tasa de transferencia sostenida. También es útil para las operaciones de copia de respaldo de disco a disco. Además, si se usa un RAID 1 o un RAID basado en división con un tamaño de bloque lo suficientemente grande se logran mejoras de rendimiento para patrones de acceso que implique múltiples lecturas simultáneas como pueden ser por ejemplo bases de datos multiusuario.

Hay que tener en cuenta que las configuraciones RAID no deben considerarse como un sistema de protección de datos ante otro tipo de desastres no relacionados con la durabilidad mecánica y física de los discos. No impedirá, por ejemplo, que un virus destruya los datos, que éstos se corrompan, que sufran la modificación o borrado accidental por parte del usuario ni que un fallo físico en otro componente del sistema afecten a los datos, etc.

RAID no mejora el rendimiento de todas las aplicaciones. Esto resulta especialmente cierto en las configuraciones típicas de escritorio. La mayoría de aplicaciones de escritorio y videojuegos hacen énfasis en la estrategia debuffering y los tiempos de búsqueda de los discos. Una mayor tasa de transferencia sostenida supone poco beneficio para los usuarios de estas aplicaciones, al ser la mayoría de los archivos a los que se accede muy pequeños. La división de discos de un RAID 0 mejora el rendimiento de transferencia lineal pero no lo demás, lo que hace que la mayoría de las aplicaciones de escritorio y juegos no muestren mejora alguna, salvo excepciones. Para estos usos, lo mejor es comprar un disco más grande y rápido, en lugar de dos discos más lentos y pequeños en una configuración RAID 0.

Los tipos de RAID frecuentemente utilizados se definen a continuación.

3.7.1.- Raid 0

Un RAID 0 distribuye los datos equitativamente entre dos o más discos (usualmente se ocupa el mismo espacio en dos o más discos) sin información de paridad que proporcione redundancia.

El RAID 0 se usa normalmente para proporcionar un alto rendimiento de lectura ya que los datos se recuperan de dos o más discos de forma paralela, aunque un mismo fichero solo está presente una vez en el conjunto.

También puede utilizarse como forma de crear un pequeño número de grandes discos virtuales a partir de un gran número de pequeños discos físicos.

Un RAID 0 puede ser creado con discos de diferentes tamaños, pero el espacio de almacenamiento añadido al conjunto estará limitado por el tamaño del disco más pequeño (por ejemplo, si un disco de 300 GB se divide con uno de 100 GB, el tamaño del conjunto resultante será sólo de 200 GB, ya que cada disco aporta 100GB).

Una buena implementación de un RAID 0 dividirá las operaciones de lectura y escritura en bloques de igual tamaño, por lo que distribuirá la información equitativamente entre los dos discos.

También es posible crear un RAID 0 con más de dos discos, si bien, la fiabilidad del conjunto será igual a la fiabilidad media de cada disco entre el número de discos del conjunto. Es decir, la fiabilidad total será, aproximadamente, inversamente proporcional al número de discos del conjunto (pues para que el conjunto falle es suficiente con que lo haga cualquiera de sus discos).

No debe confundirse RAID 0 con un Volumen Distribuido (Spanned Volume) en el cual se agregan múltiples espacios no usados de varios discos para formar un único disco virtual

3.7.2.- RAID 1 (Mirroring)

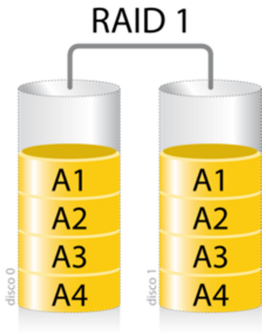
Un RAID 1 crea una copia exacta (o espejo) de un conjunto de datos en dos o más discos. Esto resulta útil cuando queremos tener más seguridad desaprovechando capacidad ya que, si perdemos un disco, tenemos el otro con la misma información.

Un RAID 1 sólo puede ser tan grande como el más pequeño de sus discos. Un RAID 1 clásico consiste en dos discos en espejo lo que incrementa exponencialmente la fiabilidad respecto a un solo disco. La probabilidad de fallo del conjunto es igual al producto de las probabilidades de fallo de cada uno de los discos (pues para que el conjunto falle es necesario que lo hagan todos sus discos).

Adicionalmente, dado que todos los datos están en dos o más discos, con hardware habitualmente independiente, el rendimiento de lectura se incrementa aproximadamente como múltiplo lineal del número de copias. Con un RAID 1 puede estar leyendo simultáneamente dos datos diferentes en dos discos diferentes, por lo que su rendimiento se duplica.

Como en el RAID 0, el tiempo medio de lectura se reduce ya que los sectores a buscar pueden dividirse entre los discos, bajando el tiempo de búsqueda y subiendo la tasa de transferencia.

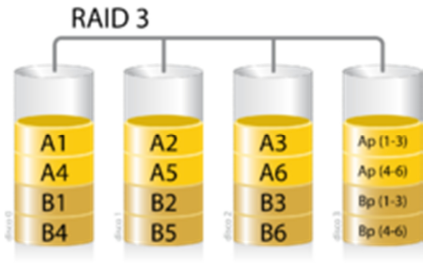
Al escribir, el conjunto se comporta como un único disco dado que los datos deben ser escritos en todos los discos del RAID 1. Por tanto, el rendimiento de escritura no mejora.



El RAID 1 tiene muchas ventajas de administración. Por ejemplo, en algunos entornos 24/7, es posible «dividir el espejo»: marcar un disco como inactivo, hacer una copia de seguridad de dicho disco y luego «reconstruir» el espejo. Esto requiere que la aplicación de gestión del conjunto soporte la recuperación de los datos del disco en el momento de la división. Este procedimiento es menos crítico que la presencia de una característica de snapshot en algunos sistemas de archivos, en la que se reserva algún espacio para los cambios, presentando una vista estática en un punto temporal dado del sistema de archivos. Alternativamente, un conjunto de discos puede ser almacenado de forma parecida a como se hace con las tradicionales cintas.

3.7.3.- RAID 3

Un RAID 3 divide los datos a nivel de bytes en lugar de a nivel de bloques . Los discos son sincronizados por la controladora para funcionar simultáneamente. Éste es el único nivel RAID original que actualmente no se usa. Permite tasas de transferencias extremadamente altas.



Teóricamente, un RAID 3 necesitaría 39 discos en un sistema informático moderno: 32 se usarían para almacenar los bits individuales que forman cada palabra y 7 se usarían para la corrección de errores.

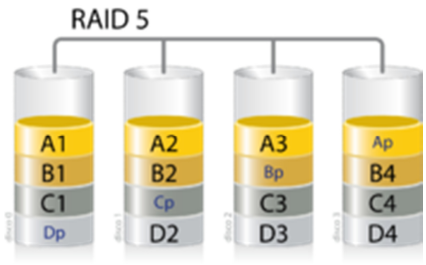
En el ejemplo del gráfico, una petición del bloque «A» formado por los bytes A1 a A6 requeriría que los tres discos de datos buscaran el comienzo (A1) y devolvieran su contenido. Una petición simultánea del bloque «B» tendría que esperar a que la anterior concluyese.

3.7.4.- RAID 5

Un RAID 5 (también llamado distribuido con paridad) es una división de datos a nivel de bloques que distribuye la información de paridad entre todos los discos miembros del conjunto.

El RAID 5 ha logrado popularidad gracias a su bajo coste de redundancia. Generalmente, el RAID 5 se implementa con soporte hardware para el cálculo de la paridad.

RAID 5 necesitará un mínimo de 3 discos para ser implementado.



En el gráfico de ejemplo anterior, una petición de lectura del bloque «A1» sería servida por el disco 0. Una petición de lectura simultánea del bloque «B1» tendría que esperar, pero una petición de lectura de «B2» podría atenderse concurrentemente ya que sería servida por el disco 1.

Cada vez que un bloque de datos se escribe en un RAID 5, se genera un bloque de paridad dentro de la misma división (stripe). Un bloque se compone a menudo de muchos sectores consecutivos de disco. Una serie de bloques (un bloque de cada uno de los discos del conjunto) recibe el nombre colectivo de división (stripe). Si otro bloque, o alguna porción de un bloque, es escrita en esa misma división, el bloque de paridad (o una parte del mismo) es recalculada y vuelta a escribir. El disco utilizado por el bloque de paridad está escalonado de una división a la siguiente, de ahí el término «bloques de paridad distribuidos». Las escrituras en un RAID 5 son costosas en términos de operaciones de disco y tráfico entre los discos y la controladora.

Los bloques de paridad no se leen en las operaciones de lectura de datos, ya que esto sería una sobrecarga innecesaria y disminuiría el rendimiento. Sin embargo, los bloques de paridad se leen cuando la lectura de un sector de datos provoca un error de CRC. En este caso, el sector en la misma posición relativa dentro de cada uno de los bloques de datos restantes en la división y dentro del bloque de paridad en la división se utiliza para reconstruir el sector erróneo. El error CRC se oculta así al resto del sistema. De la misma forma, si falla un disco del conjunto, los bloques de paridad de los restantes discos son combinados matemáticamente con los bloques de datos de los restantes discos para reconstruir los datos del disco que ha fallado «al vuelo».

Lo anterior se denomina a veces Modo Interino de Recuperación de Datos (Interim Data Recovery Mode). El sistema sabe que un disco ha fallado, pero sólo con el fin de que el sistema operativo pueda notificar al administrador que una unidad necesita ser reemplazada: las aplicaciones en ejecución siguen funcionando ajenas al fallo. Las lecturas y escrituras continúan normalmente en el conjunto de discos, aunque con alguna degradación de rendimiento. La diferencia entre el RAID 4 y el RAID 5 es que, en el Modo Interno de Recuperación de Datos, el RAID 5 puede ser ligeramente más rápido, debido a que, cuando el CRC y la

paridad están en el disco que falló, los cálculos no tienen que realizarse, mientras que en el RAID 4, si uno de los discos de datos falla, los cálculos tienen que ser realizados en cada acceso.

El fallo de un segundo disco provoca la pérdida completa de los datos.

El número máximo de discos en un grupo de redundancia RAID 5 es teóricamente ilimitado, pero en la práctica es común limitar el número de unidades. Los inconvenientes de usar grupos de redundancia mayores son una mayor probabilidad de fallo simultáneo de dos discos, un mayor tiempo de reconstrucción y una mayor probabilidad de hallar un sector irrecuperable durante una reconstrucción. A medida que el número de discos en un conjunto RAID 5 crece, el MTBF (tiempo medio entre fallos) puede ser más bajo que el de un único disco. Esto sucede cuando la probabilidad de que falle un segundo disco en los N-1 discos restantes de un conjunto en el que ha fallado un disco en el tiempo necesario para detectar, reemplazar y recrear dicho disco es mayor que la probabilidad de fallo de un único disco. Una alternativa que proporciona una protección de paridad dual, permitiendo así mayor número de discos por grupo, es el RAID 6.

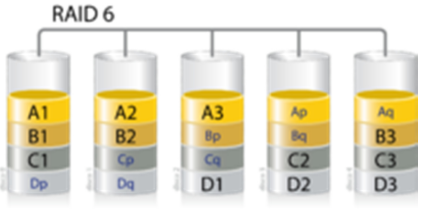
Algunos vendedores RAID evitan montar discos de los mismos lotes en un grupo de redundancia para minimizar la probabilidad de fallos simultáneos al principio y el final de su vida útil.

Las implementaciones RAID 5 presentan un rendimiento malo cuando se someten a cargas de trabajo que incluyen muchas escrituras más pequeñas que el tamaño de una división (stripe). Esto se debe a que la paridad debe ser actualizada para cada escritura, lo que exige realizar secuencias de lectura, modificación y escritura tanto para el bloque de datos como para el de paridad. Implementaciones más complejas incluyen a menudo cachés de escritura no volátiles para reducir este problema de rendimiento.

En el caso de un fallo del sistema cuando hay escrituras activas, la paridad de una división (stripe) puede quedar en un estado inconsistente con los datos. Si esto no se detecta y repara antes de que un disco o bloque falle, pueden perderse datos debido a que se usará una paridad incorrecta para reconstruir el bloque perdido en dicha división. Esta potencial vulnerabilidad se conoce a veces como «agujero de escritura». Son comunes el uso de caché no volátiles y otras técnicas para reducir la probabilidad de ocurrencia de esta vulnerabilidad.

3.7.5.- RAID 6

Un RAID 6 amplía el nivel RAID 5 añadiendo otro bloque de paridad, por lo que divide los datos a nivel de bloques y distribuye los dos bloques de paridad entre todos los miembros del conjunto. El RAID 6 no era uno de los niveles RAID originales.



Al igual que en el RAID 5, en el RAID 6 la paridad se distribuye en divisiones (stripes) con los bloques de paridad en un lugar diferente en cada división.

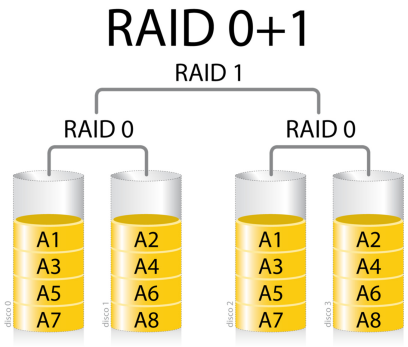
El RAID 6 es ineficiente cuando se usa un pequeño número de discos, pero a medida que el conjunto crece y se dispone de más discos la pérdida en capacidad de almacenamiento se hace menos importante, creciendo al mismo tiempo la probabilidad de que dos discos fallen simultáneamente. El RAID 6 proporciona protección contra fallos dobles de discos y contra fallos cuando se está reconstruyendo un disco. En caso de que sólo tengamos un conjunto puede ser más adecuado que usar un RAID 5 con un disco de reserva (hot spare).

La capacidad de datos de un conjunto RAID 6 es n-2, siendo n el número total de discos del conjunto.

Un RAID 6 no penaliza el rendimiento de las operaciones de lectura, pero sí el de las de escritura debido al proceso que exigen los cálculos adicionales de paridad. Esta penalización puede minimizarse agrupando las escrituras en el menor número posible de divisiones (stripes), lo que puede lograrse mediante el uso de un sistema de archivos WAFL.

3.7.6.- RAID 0+1

Un RAID 0+1 (también llamado RAID 01) es un RAID usado para replicar y compartir datos entre varios discos. La diferencia entre un RAID 0+1 y un RAID 1+0 es la localización de cada nivel RAID dentro del conjunto final. Un RAID 0+1 es un espejo de divisiones.



Como puede verse en el diagrama, primero se crean dos conjuntos RAID 0 (dividiendo los datos en discos) y luego, sobre los anteriores, se crea un conjunto RAID 1 realizando un espejo de los anteriores.

La ventaja de un RAID 0+1 es que cuando un disco duro falla los datos perdidos pueden ser copiados del otro conjunto de nivel 0 para reconstruir el conjunto global. Sin embargo, al añadir un disco duro adicional en una división es obligatorio añadir otro al de la otra división para equilibrar el tamaño del conjunto.

El RAID 0+1 no es tan robusto como un RAID 1+0 que veremos en el punto siguiente no pudiendo tolerar dos fallos simultáneos de discos salvo que sean en la misma división. Es decir, cuando un disco falla, la otra división se convierte en un punto de fallo único. Además, cuando se sustituye el disco que falló se necesita que todos los discos del conjunto participen en la reconstrucción de los datos.

Con la cada vez mayor capacidad de las unidades de discos (liderada por las unidades serial ATA), el riesgo de fallo de los discos es cada vez mayor. Además, las tecnologías de corrección de errores de bit no han sido capaces de mantener el ritmo de rápido incremento de las capacidades de los discos, provocando un mayor riesgo de hallar errores físicos irrecuperables.

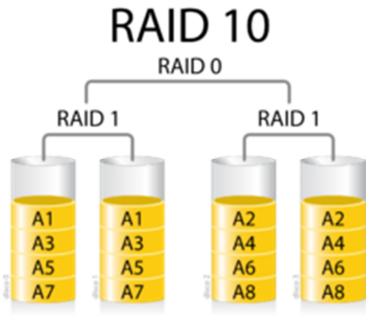
Debido a esto cada vez tiene mayores riesgos el RAID 0+1 y su vulnerabilidad ante los fallos dobles simultáneos)

3.7.7.- RAID 1+0

Un RAID 1+0, a veces llamado RAID 10, es parecido a un RAID 0+1 con la excepción de que los niveles RAID que lo forman se invierte: el RAID 10 es una división de espejos.

En cada división RAID 1 pueden fallar todos los discos salvo uno sin que se pierdan datos. Sin embargo, si los discos que han fallado no se reemplazan, el restante pasa a ser un punto único de fallo para todo el conjunto. Si ese disco falla entonces, se perderán todos los datos del conjunto completo.

Como en el caso del RAID 0+1, si un disco que ha fallado no se reemplaza, entonces un solo error de medio irrecuperable que ocurra en el disco espejado resultaría en pérdida de datos.



Debido a estos mayores riesgos del RAID 1+0, muchos entornos empresariales críticos están empezando a evaluar configuraciones RAID más tolerantes a fallos que añaden un mecanismo de paridad subyacente. Entre los más prometedores están los enfoques híbridos como el RAID 0+1+5 (espejo sobre paridad única) o RAID 0+1+6 (espejo sobre paridad dual).

El RAID 10 es a menudo la mejor elección para bases de datos de altas prestaciones, debido a que la ausencia de cálculos de paridad proporciona mayor velocidad de escritura.

3.7.8.- Comparación de los diferentes niveles de raid:

Como resumen, la comparativa entre los diferentes niveles RAID se puede resumir en el siguiente cuadro:

RAID	MIN. DISKS	STORAGE EFFICIENCY %	COST	READ PERFORMANCE	WRITE PERFORMANCE	WRITE PENALTY
0	2	100	Low	Very good for both random and sequential read	Very good	No
1	2	50	High	Good. Better than a single disk.	Good. Slower than a single disk, as every write must be committed to all disks.	Moderate
3	3	(n-1)*100/n where n= number of disks	Moderate	Good for random reads and very good for sequential reads.	Poor to fair for small random writes. Good for large, sequential writes.	High
4	3	(n-1)*100/n where n= number of disks	Moderate	Very good for random reads. Good to very good for sequential writes.	Poor to fair for random writes. Fair to good for sequential writes.	High
5	3	(n-1)*100/n where n= number of disks	Moderate	Very good for random reads. Good for sequential reads	Fair for random writes. Slower due to parity overhead. Fair to good for sequential writes.	High
6	4	(n-2)*100/n where n= number of disks	Moderate but more than RAID 5	Very good for random reads. Good for sequential reads.	Good for small, random writes (has write penalty).	Very High
1+0 and 0+1	4	50	High	Very good	Good	Moderate

CAPÍTULO 4.- Industria de los sistemas de almacenamiento

4.1.- EMC

Comenzó sus actividades el 23 de agosto de 1979, fundada por Richard (Dick) Egan (ex gerente de Intel) y Roger Marino, la E y la M en el nombre de la compañía (EMC no adoptó EMC² refiriéndose a la famosa ecuación E=mc² de Einstein). La primera C se refiere a un tercer integrante que dejó la compañía luego de fundada y la segunda se refiere a Corporation.

Inicialmente se dedicaban a fabricar placas de memoria, expandiéndose luego a manejadores de disco.

El 4 de abril de 1986 comienza a cotizar en bolsa e integra el NASDAQ. y el 22 de marzo de 1988 integra el NYSE.

Con el trabajo de Moshe Yanai (que trabaja actualmente para XIV Storage), creció hasta convertirse en la más importante empresa en innovaciones y plataformas de almacenamiento. Joseph Tucci es el actual CEO desde 2001, remplazando a Michael Ruettgers que continuó como chairman hasta el 2006.

La compañía se transformó de una compañía de hardware a una compañía mezcla de hardware, software y servicios profesionales. El gran impulso para el futuro es el desarrollo de productos para virtualización que incluyen VMware, Invista, and Rainfinity.

El 12 de octubre de 2015 se anuncia que la compañía Dell Inc. compra EMC Corporation por 67 mil millones de dólares, convirtiéndose en el mayor acuerdo de tecnología de todos los tiempos.1

- Dentro de sus hitos existen infinidad de innovaciones tanto en memorias, como en sistemas de almacenamiento, en sus inicios enfocados a mainframe. La misma tiene múltiples alianzas con Oracle, Microsoft, Dell y SAP.
- 23 de agosto de 1979 se funda la compañía
- 1990 lanza al mercado el Symmetrix
- 1992 se lanza al mercado la serie 5500 de Symmetrix, el cual garantiza una operación 7 X 24.
- 1994 anuncia el primer arreglo mundial de discos que superan 1 TB; con el modelo 5500-9.
- 1999 adquiere Data General, y con ella CLARiiON.
- 2000 une NAS y SAN en una red unificada.
- 2003 su solución de almacenamiento CLARiiON, es la primera en usar discos de bajo precio con tecnología SATA y Canal de Fibra.
- En julio de 2006 EMC abre una oficina de desarrollo en Shanghai, China, para facilitar el ingreso en el mercado Chino.
- El 7 de julio de 2007, anuncia la inversión de US$ 160 millones en Singapur, para la construcción de un nuevo laboratorio de 15.000 pies, el cual comenzará a operar este año.
- El 12 de noviembre de 2007 se asocia con NetQoS, para proveer la primera solución integrada de monitorización.

- El 20 de diciembre de 2013 la agencia de noticias Reuters publica que las revelaciones de Edward Snowden informen un posible acuerdo económico secreto entre la empresa RSA, filial de EMC Corp, y la NSA para que el software criptográfico de la empresa (Bsafe) use como generador de números aleatorios por defecto el algoritmo Dual_EC_DRBG a sabiendas de que tenía debilidades. De esta forma la NSA se aseguraría una puerta trasera en dicho software.
- Las adquisiciones y directivas ayudaron al crecimiento de EMC hasta convertirla en una de las mayores empresas en el desarrollo, y construcción de equipos de almacenamiento de datos en el mundo.

4.2.- HP

Hewlett-Packard (NYSE: HPQ), más conocida como HP, es una empresa estadounidense, de las mayores empresas de tecnologías de la información del mundo, con sede en Palo Alto, California. Fabricaba y comercializaba hardware y software además de brindar servicios de asistencia relacionados con la informática. La compañía fue fundada en 1939 por William Hewlett y David Packard, y se dedicaba a la fabricación de instrumentos de medida electrónica y de laboratorio.

El 6 de octubre de 2014, Hewlett-Packard anuncio su división en dos firmas que cotizarían de manera separada en el mercado de valores, con lo que su negocio de computadoras e impresoras operaría independiente de su unidad de servicios y equipos corporativos.

El 1 de noviembre se hizo efectiva su división en dos empresas: HP Inc., dedicada a las impresoras y las computadoras personales, y Hewlett Packard Enterprise, dedicada a los servidores, equipos de almacenamiento y redes, programas de cómputo y servicios para terceras empresas. En el proceso de separación se planea despedir a 5000 empleados de la firma.

4.3.- IBM

International Business Machines Corp. (IBM) (NYSE: IBM) es una reconocida empresa multinacional estadounidense de tecnología y consultoría con sede en Armonk, Nueva York. IBM fabrica y comercializa hardware y software para computadoras, y ofrece servicios de infraestructura, alojamiento de Internet, y consultoría en una amplia gama de áreas relacionadas con la informática, desde computadoras centrales hasta nanotecnología.

La empresa fue fundada en 1911 como Computing Tabulating Recording Corporation, el resultado de la fusión de cuatro empresas: Tabulating Machine Company, International Time Recording Company, Computing Scale Corporation, y Bundy Manufacturing Company.3 4 CTR adoptó el nombre International Business Machines en 1924, utilizando un nombre previamente designado a un filial de CTR en Canadá, y posteriormente en América del Sur.

En 2011, la revista Fortune clasificó IBM como la empresa número 18 en los Estados Unidos en tamaño, y la empresa número 7 en beneficios.

Globalmente, la empresa fue clasificada como la empresa número 31 en tamaño por Forbes en 2011.7 8 Por el número de empleados (más de 425.000, quienes se denominan como "IBMers") es la segunda

empresa más grande del mundo solo superada por Walmart (en más de 200 países, con ocupaciones incluyendo científicos, ingenieros, consultores y profesionales de ventas).

IBM alberga más patentes que ninguna otra empresa de tecnología de Estados Unidos, y tiene nueve laboratorios de investigación.10 Sus empleados han recibido cinco Premios Nobel, cuatro Premios Turing, nueve National Medals of Technology y cinco National Medals of Science.11 Las invenciones famosas de IBM incluyen el cajero automático, el disquete, el disco duro, la banda magnética, el modelo relacional, el Universal Product Code, el "financial swap," el sistema de reservas aéreas SABRE, DRAM y el sistema de inteligencia artificial Watson.

4.4.- Netapp

NetApp fue fundada en 1992 por David Hitz, James Lau, y Michael Malcolm. [En ese momento, su principal competidor era Auspex Systems.

En 1994, NetApp recibido fondos de capital de riesgo de Sequoia Capital y tuvo su oferta pública inicial en 1995. NetApp prosperó en los años de la burbuja de Internet de mediados de la década de 1990 a 2001, durante el cual la compañía creció mil millones de dólares en ingresos anuales.

Después de la explosión de la burbuja, los ingresos de NetApp disminuyeron rápidamente a $ 800 millones en el año fiscal 2002. Desde entonces, la facturación de la compañía ha aumentado constantemente.

En 2006, NetApp vendió la línea de productos NetCache de Blue Coat Systems. En 2014, NetApp adquirió la línea de Riverbed Technology SteelStore de copia de seguridad y de protección de datos de productos, [10], que más tarde rebautizado como AltaVault. El 1 de junio de 2015, Tom Georgens renunció como CEO y fue sustituido por George Kurian. En diciembre de 2015, NetApp adquirió almacenamiento flash SolidFire proveedor por 870 millones de dólares.

CAPÍTULO 5.- Protección de los sistemas de almacenamiento

5.1.- RTO y RPO en sistemas de almacenamiento

Las configuraciones tradicionales de almacenamiento proponían un sistema activo/pasivo en el cual, mediante el uso de sistemas de réplicas, se podía disponer de una alta disponiblidad entre sites separados entre ellos varios kilómetros.

La limitación de distancia entre sites viene dada por la limitación física de la transmisión de la información por el medio físico elegido y la forma de réplica configurada.

Para la configuración adecuada de las réplicas es necesario disponer de un centro de respaldo (denominado comúnmente CPD o Centro de Proceso de Datos) diseñado bajo los mismos principios del CPD principal pero añadiendo algunas consideraciones más.

En primer lugar, debe elegirse una localización totalmente distinta a la del CPD principal con el objeto de que no se vean ambos afectados simultáneamente por la misma contingencia. Es habitual situarlos entre 20 y 40 kilómetros del CPD principal. La distancia está limitada por las necesidades de telecomunicaciones entre ambos centros.

En segundo lugar, el equipamiento electrónico e informático del centro de respaldo debe ser absolutamente compatible con el existente en el CPD principal. Esto no implica que el equipamiento deba ser *exactamente* igual ya que, normalmente, no todos los procesos del CPD principal son críticos.

Por otra parte, tampoco se requiere el mismo nivel de servicio en caso de emergencia. En consecuencia, es posible utilizar hardware menos potente

En tercer lugar, el equipamiento software debe ser idéntico al existente en el CPD principal. Esto implica exactamente las mismas versiones y parches del software de base y de las aplicaciones corporativas que estén en producción en el CPD principal. De otra manera, no se podría garantizar totalmente la continuidad de operación.

Por último, pero no menos importante, es necesario contar con una réplica de los mismos datos con los que se trabaja en el CPD original.

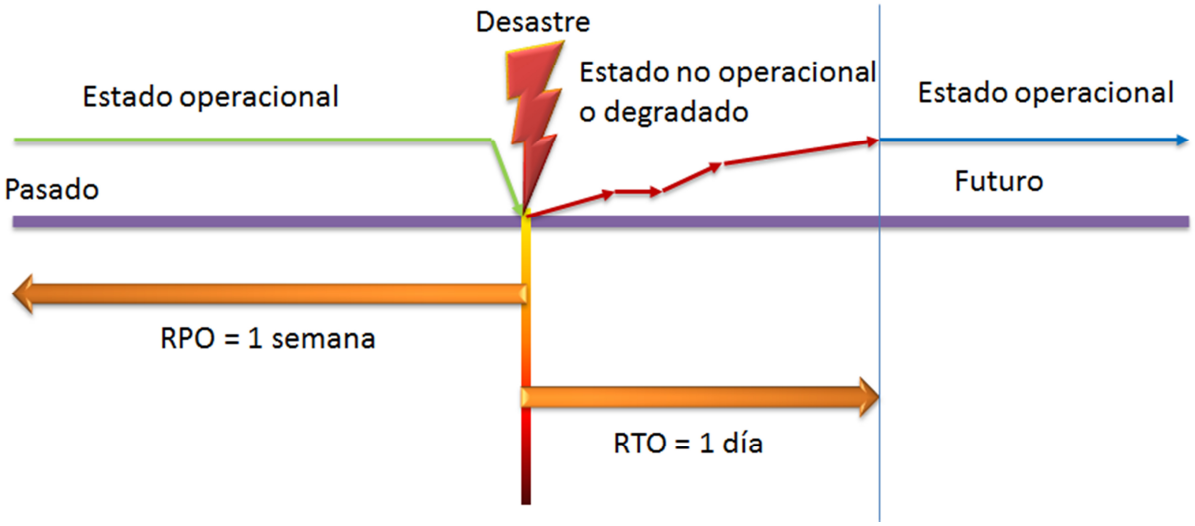
Existen dos políticas o aproximaciones a este problema:

- Copia síncrona de datos: Se asegura que todo dato escrito en el CPD principal también se escribe en el centro de respaldo antes de continuar con cualquier otra operación.
- Copia asíncrona de datos: No se asegura que todos los datos escritos en el CPD principal se escriban inmediatamente en el centro de respaldo, por lo que puede existir un desfase temporal entre unos y otros.

La copia asíncrona puede tener lugar fuera de línea. En este caso, el centro de respaldo utiliza la última copia de seguridad existente del CPD principal. Esto lleva a la pérdida de los datos de operaciones de varios minutos hasta días. Esta opción es viable para negocios no demasiado críticos, donde es más importante la continuidad del negocio que la pérdida de datos. Por ejemplo, en cadenas de supermercados o pequeños negocios. No obstante, es inviable en negocios como la banca, donde es impensable la pérdida de una sola transacción económica.

Tanto para la copia síncrona como asíncrona, es necesaria una extensión de la red de almacenamiento entre ambos centros. Es decir, un enlace de telecomunicaciones entre el CPD y el centro de respaldo. En caso de copia síncrona es imprescindible que dicho enlace goce de baja latencia. Motivo por el que se suele emplear un enlace de fibra óptica, que limita la distancia máxima a decenas de kilómetros

La decisión técnica en el uso de sistemas de replicación síncronas o asíncronas, así como en la elección del sistema de almacenamiento que más se adapta a las necesidades del negocio se toman en base a dos conceptos: RTO y RPO.



5.1.1.- RPO (recovery point objective)

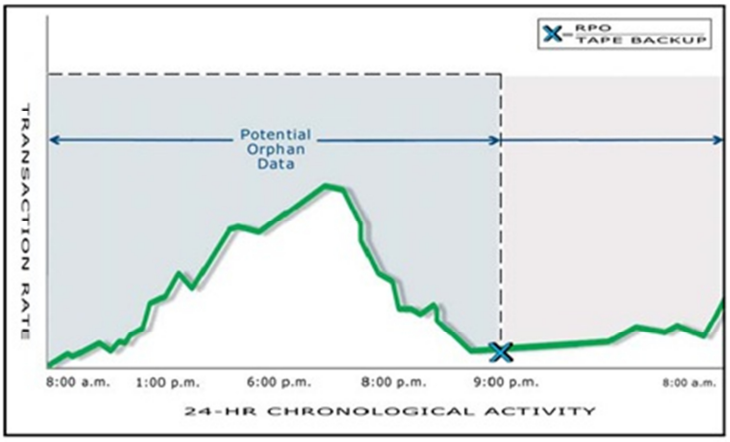
RPO se refiere al volumen de datos en riesgo de pérdida que la organización considera tolerable. Esto es, las transacciones de cuánto tiempo estamos dispuestos a perder, o a tener que reintroducir al sistema.

La respuesta va a depender del volumen de transacciones por unidad de tiempo y de los mecanismos de backup, pero siempre aumenta el volumen de datos 'huérfanos' a medida que pasa el tiempo desde la última copia de seguridad.

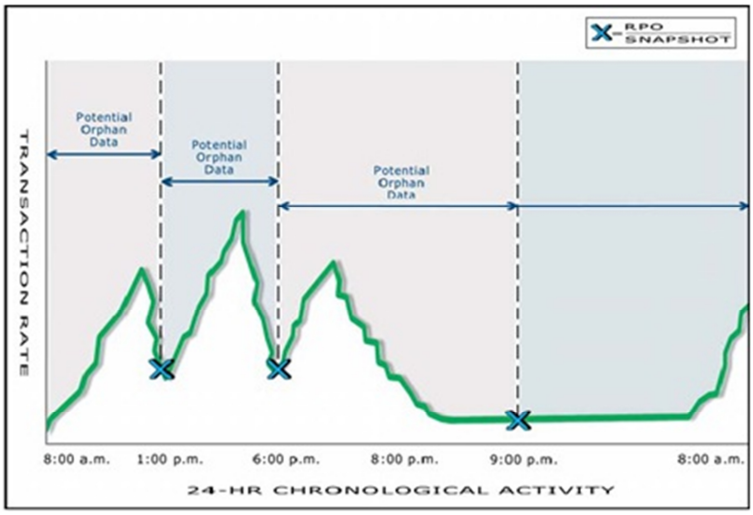
El RPO determina el objetivo de posible pérdida máxima de datos introducidos desde el último backup, hasta la caída del sistema, y no depende del tiempo de recuperación. La casuística es amplísima, y aquí se ilustran algunos casos:

Dependiente del backup:

- Caso 1: Backup diario en cinta a las 09:00 PM



- Caso 2: Varios snapshots durante el día



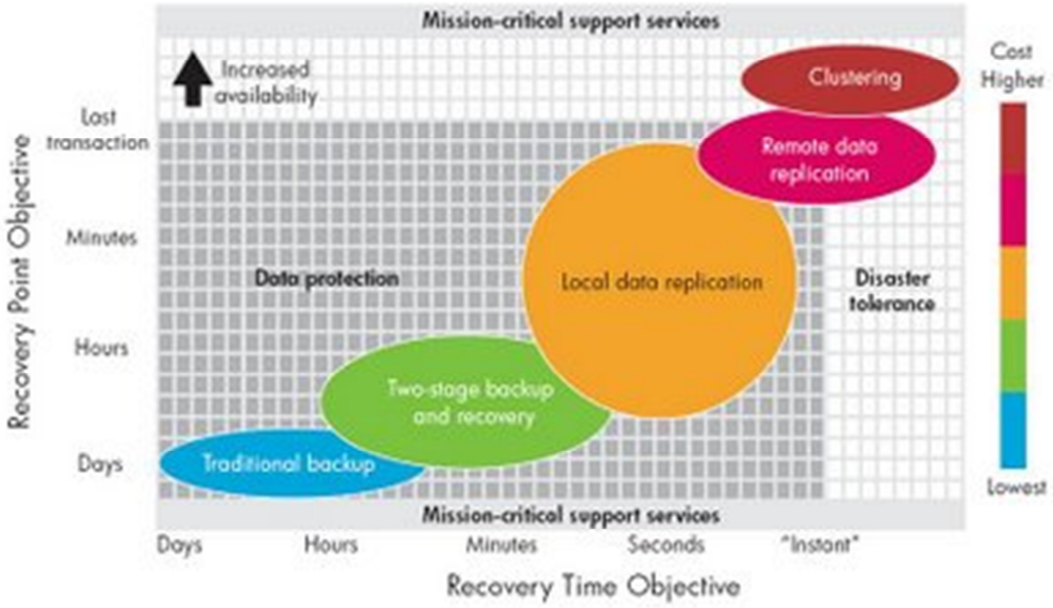
Dependiendo de la naturaleza del negocio

- Caso 3: La gestión de las transacciones bursátiles no puede parar, no se puede perder ni demorar ni una sola transacción, su RPO es 0.
- Caso 4: Un almacén robotizado se parará si no tiene sistema, y con mucho movimiento estará en situación de colapso en pocos minutos

5.1.2.- RTO (recovery time objective)

Expresa el tiempo durante el cual una organización puede tolerar la falta de funcionamiento de sus aplicaciones y la caída de nivel de servicio asociada sin afectar a la continuidad del negocio.

La respuesta dependerá de la criticidad de cada aplicación. Por ejemplo, no será lo mismo la aplicación que da servicio a las cajas en una gran superficie que la aplicación para el cálculo de la nómina, que se ejecuta una vez al mes.



Como se puede observar en el diagrama anterior, el mayor coste económico a la hora de decidir qué sistema implantar es aquel en el tendremos un RTO=0 y un RPO=0 o tendiendo a 0.

5.2.- Costes de los tiempos muertos

Resulta relativamente sencillo calcular el coste que le supone a una organización la pérdida de servicio por una contingencia en su centro de respaldo.

Como método estándar, se suelen calcular en base a 3 conceptos: La pérdida de ingresos, los costes de personal y los “costes intangibles”

Desglosando cada uno de los conceptos podemos definir:

- Pérdida de ingresos: Podemos usar una fórmula simple para calcular el ingreso por una hora de funcionamiento del negocio, expresada de forma:

Ingresos por hora = Ingreso anual total / Horas hábiles de negocio al año

ingresos anual total -> calculados en Euros por hora

horas hábiles de negocio al año -> calculadas con la siguiente fórmula:

Horas de negocio al año = Horas de negocio al día * días de negocio al mes * meses de negocio al año

Para obtener el valor de la pérdida calcularemos:

Pérdida de ingresos = ingresos por hora * número de horas de tiempo muerto.

- Cálculo del coste de personal: Se deben tener en cuenta este tipo de costes derivados de tener un sistema inaccesible

Pérdida laboral = número de personas * % de horas afectado * número de horas * € /hora/empleado

dónde €/hora/empleado es el coste por hora de empleado, definido cómo:

€/hora/empleado = paga por hora + beneficios sociales + imputación gastos fijos

- Cálculo de costes intangibles

Por más intangibles que sean, estos costes son algo que tiene que ser tomado en consideración, porque ponen de manifiesto algunos de los efectos a largo plazo de los fallos de disponibilidad de sistemas. Alguna de las cosas que hay que tener en cuenta:

4. Pérdida de reputación: ¿Quizás sus clientes no sean tan fieles, después de una incidencia que impidió a su empresa satisfacerles, durante un período de tiempo? ¿Es esta la oportunidad para ellos, para “salir de compras por allí”?
5. El efecto secundario sobre los clientes de sus clientes. ¿Hay otras empresas o personas, aparte de sus clientes primarios que puedan ser afectadas?
6. El ánimo dentro de la empresa. Si los empleados no pueden cumplir con su trabajo, debido a fallos del sistema, su ánimo afecta su modo de trabajar y prestar servicios.

Aplicando las premisas de la teoría anterior aplicadas a un negocio que funciona de 8 a.m. a 8 p.m., de lunes a sábado, todas las semanas del año. Los ingresos por ventas del año pasado fueron de 36.060.726,26 €. Si procesamos estos datos con nuestra ecuación obtenemos:

Ingresos por hora = 11.382,81 € = 36.060.726,26 € / (12 horas al día * 22 días al mes * 12 meses al año) à = 36.060.726,26 € / 3.168 horas

El ingreso por hora asciende a 11.382,81 €. De acuerdo con éste cálculo, una incidencia de 12 horas de tiempo muerto, implica una pérdida de ingresos que asciende a 136.593,72 € (12 horas por 11.382,81 €)

El considerar que el negocio no genera ingresos, mientras el sistema no esté disponible, es indudablemente una suposición importante. Pero como mínimo, este enfoque nos da un punto de partida y unos datos estadísticos, para comprender el valor de cada hora disponible de la organización.

Por ejemplo, supongamos que en la empresa, el coste total promedio por empleado por hora es de 15€. Si 50 empleados son afectados en un 50% durante las primeras 2 horas del fallo del sistema, 100 personas se ven afectadas en un 75% por las próximas 2 horas y 200 personas pierden todo su tiempo (100%), cuando el fallo dura más de 4 horas, un tiempo muerto de 12 horas costaría a la empresa euros. 27.000€.

Nº de Personas		% de horas afectadas		Horas fallo ordenador		Euros/hora empleado		Coste de personal
50	x	50%	x	2	x	15 €	=	750 €
100	x	75%	x	2	x	15 €	=	2.250 €
200	x	100%	x	8	x	15 €	=	24.000 €
								27.000 €

La figura anterior muestra los cálculos correspondientes. Dividiendo este coste total de personal, entre el número de horas que duró la incidencia (12), obtenemos el coste promedio por hora: 2.250€. Esto quiere decir que cada hora de fallo de ordenador le cuesta a la empresa un promedio de 2.250€ en personal.

El resumen de costes quedaría de la siguiente manera:

	€/hora	Por 36 horas
Pérdida de ingresos	10.820 €	389.500 €
Costes personal	2.250 €	81.130 €
Costes intangibles	4.250 €	151.456 €
Coste total del fallo	17.300 €	622.047 €

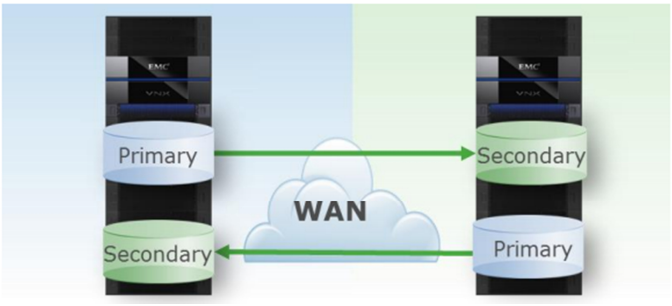
5.3.- Réplicas

Una de las decisiones de continuidad de negocio más importantes a nivel de almacenamiento de datos es la utilización de réplicas.

En los sistemas de almacenamiento más modernos la configuración de réplicas es relativamente sencilla de configurar, gestionar y administrar.

Las réplicas están formadas por los siguientes elementos:

- Primary image: Origen de la réplica, corresponde a la LUN, volumen, File System… que está dando servicio y cuyo estado es de lectura/escritura (RW) para cualquier host que tenga acceso a ella.
- Secondary image: Destino de la réplica, corresponde a la LUN, volumen, File System… que está en la cabina remota en modo NA (Not available).



- Consistency Group: Agrupación de réplicas por entorno operativo que permiten respetar el orden de llegada de las IOPS a todos los volúmenes pertenecientes al grupo.

A la hora de configurar réplicas, atendiendo al nivel de criticidad por servicio, existen 2 tipos de replicaciones:

5.2.1.- Replicación síncrona

Escribe los datos en el site primario y secundario a la vez, de manera síncrona. Esto provoca una latencia extra a la hora de escribir ya que cualquier escritura sigue el siguiente proceso:

1. El host lanza una escritura
2. La escritura se escribe en la cabina local
3. La escritura se copia de manera síncrona a la cabina remota
4. La cabina remota devuelve el ACK a la cabina principal
5. La cabina local devuelve el ACK al host

El principal beneficio de la replicación síncrona es que siempre se mantiene una copia consistente entre origen y destino de forma que ante cualquier contingencia nos aseguramos de que tendremos una copia siempre actualizada. Nunca tendremos pérdida de datos a nivel de host y hablaremos, en este caso, de un RPO = 0.

Por regla general, el RTT (Round trip time) en este tipo de réplicas no debe superar los 10ms para poder asegurar la consistencia de la copia y reducir al máximo la latencia añadida por la sincronía. Si este RTT es demasiado alto afectará al volumen de producción que está dando servicio.

5.2.2.- Replicación asíncrona

Cada escritura que llega a la cabina local (site A) es escrita en caché y el ACK devuelto al host. Como podemos ver, en este caso no encontramos latencia extra al no tener que escribir en el site remoto. En la configuración de la réplica es necesario establecer un tiempo de replicación que será dependiente del RPO establecido según las necesidades del negocio.

En este tipo de configuración el ACK del host se devuelve inmediatamente una vez la escritura llega a la cabina local.

Las réplicas asíncronas hacen uso de los snapshots para saber qué datos deben copiar desde la última replicación.

Si se establece un tiempo de replicación de, por ejemplo, 10 minutos a partir de las 9.00h, el proceso realizado será:

1. Creación de snapshot a las 9.00h
2. Sincronización full entre origen y destino
3. A las 9.10h, creación de snapshot.
4. Copia de datos diferentes (incrementales) entre el snapshot de las 9.10h y 9.00h
5. Borrado de snapshot de las 9.00h.

El RTT en este tipo de réplicas no debería exceder los 200ms.

5.2.3.- Parámetros decisores del tipo de replicación

La decisión a la hora de establecer qué tipo de réplica configurar entre sites depende de:

- Medio físico utilizado: El uso de fibra limita la distancia máxima de separación entre sites. Es por ello que si la distancia entre ellos es una limitación, el uso de una réplica asíncrona a través de TCP/IP puede ser la solución.
- Protocolo utilizado: Fiber Channel vs iSCSI
- Tecnología usada: No todos los sistemas de almacenamiento permiten la replicación síncrona porque el producto puede no estar pensado para un disaster recovery.
- Limitación en el ancho de banda: Si el BW entre sites es una limitación, las réplicas asíncronas por normal general ocupan menos el canal que las réplicas síncronas.

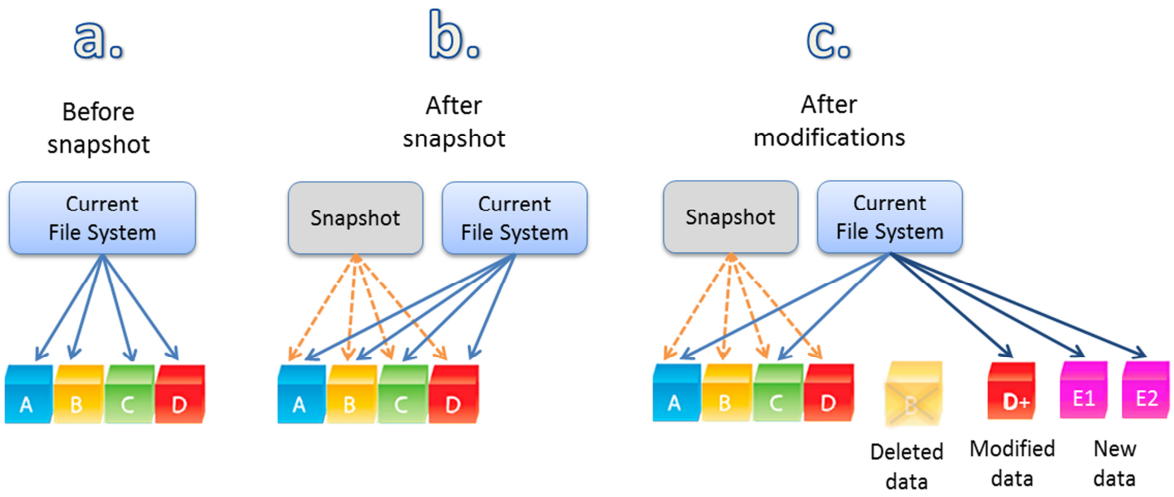
5.4.- Snapshots

Un snapshot es una instantánea de un disco, volumen o LUN en un determinado instante. Están basadas en punteros y requieren de un área específica para poder guardar los cambios que se realizan.

A diferencia de las réplicas, los snapshots solo se realizan a nivel de local dentro de una misma cabina de almacenamiento. Es por ello que no suele considerarse como un método de disaster recovery aunque puede ayudar a la hora de recuperar datos.

Cada fabricante implementa su propio sistema de snaps pero su funcionamiento se suele basar en el Copy on First Write (COFW). Cuando se crea un snapshot de un volumen ocurre lo siguiente:

1. Creación del snapshot
2. Se crea un nuevo volumen que es el resultado de los punteros que apuntan a los bloques del volumen origen
3. Si el volumen origen se modifica, el contenido del bloque es copiado dentro del área reservada y se cambia el puntero del snapshot hacia ese otro bloque, de manera que el snapshot siempre tendrá una “copia” de los datos originales.

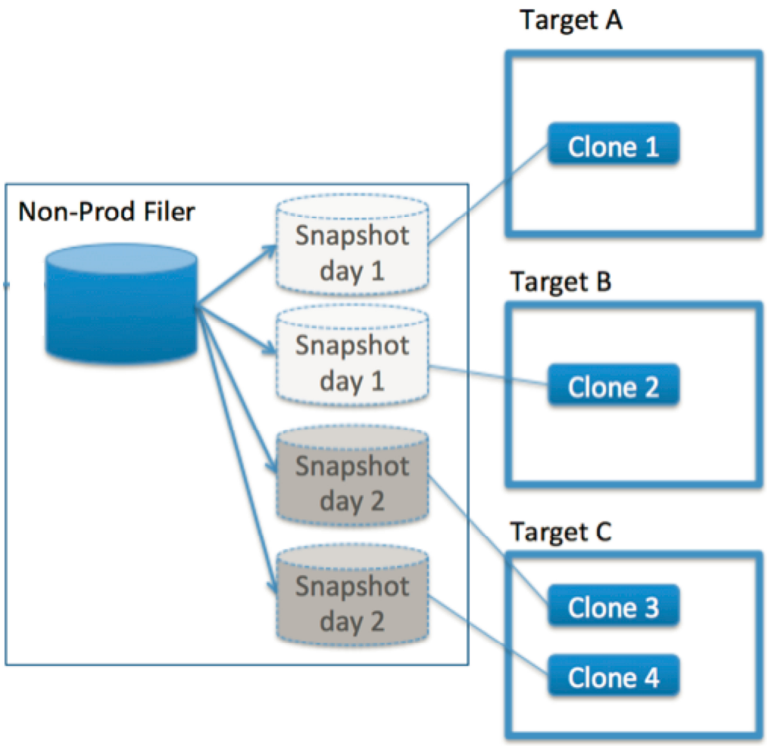


A pesar de que los snapshots se basan en punteros, requiere de un área específica que se encargará de almacenar los cambios del volumen origen. Y este área ocupa espacio, almacenamiento, lo que supone un coste económico. Por regla general, se suele reservar en torno a un 20% del espacio total que se tiene pensado proteger.

Los snapshots se pueden crear, refrescar, eliminar o restaurar. Con la utilización de sistemas Windows, por ejemplo, es posible restaurar ficheros haciendo uso de la herramienta “versiones anteriores”.

Los snapshots difieren de los clones en el espacio utilizado para su creación. Si bien ambos tienen sentido en un ámbito local, los clones necesitan el 100% del espacio provisionado del volumen origen a partir del cual están creados.

Es posible crear snapshots según el tiempo establecido y, una vez creado, consolidarlo en forma de clon, de manera que podamos tener una copia full del volumen origen en el momento deseado:



5.5.- Backup

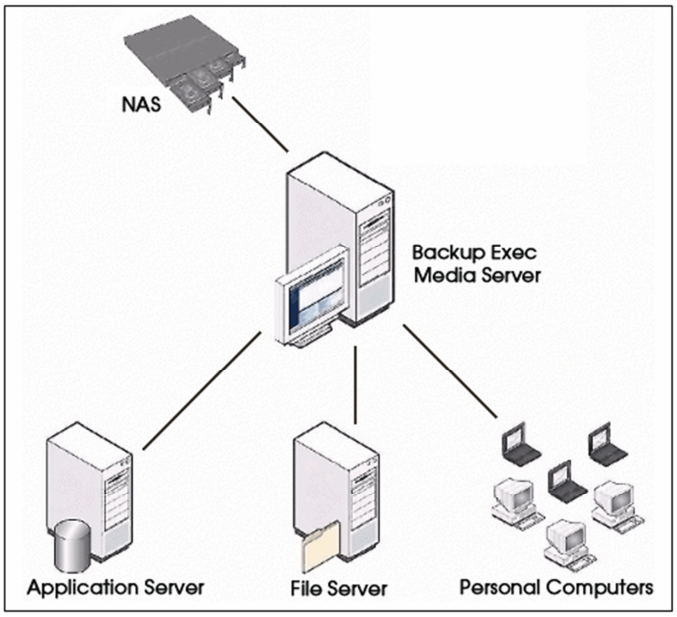
Una copia de seguridad, copia de respaldo o también llamado backup es una copia de los datos originales que se realiza con el fin de disponer de un medio para recuperarlos en caso de su pérdida. Garantizan la integridad y seguridad.

Las copias de seguridad son útiles ante distintos eventos y usos:

- Recuperar los sistemas informáticos y los datos de una catástrofe informática, natural o ataque
- Restaurar una pequeña cantidad de archivos que pueden haberse eliminado accidentalmente, corrompido, infectado por un virus informático u otras causas
- Guardar información histórica de forma más económica que los discos duros permitiendo, además, el traslado a ubicaciones distintas de la de los datos originales.

El proceso de copia de seguridad se complementa con otro conocido como restauración de los datos (en inglés restore), que es la acción de leer y grabar en la ubicación original u otra alternativa los datos requeridos.

Ya que los sistemas de respaldo contienen por lo menos una copia de todos los datos que vale la pena salvar deben tenerse en cuenta los requerimientos de almacenamiento. La organización del espacio de almacenamiento y la administración del proceso de efectuar la copia de seguridad son tareas complicadas.



Antes de que los datos sean enviados a su lugar de almacenamiento se lo debe seleccionar, extraer y manipular.

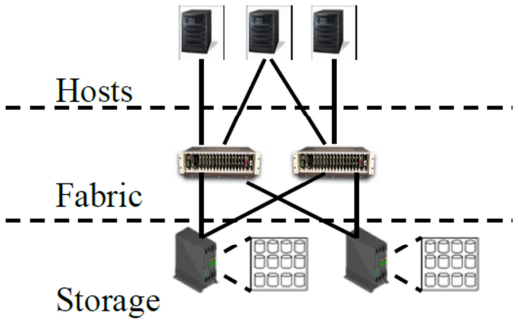
Se han desarrollado muchas técnicas diferentes para optimizar el procedimiento de efectuar los backups. Estos procedimientos incluyen, entre otros, optimizaciones para trabajar con archivos abiertos y fuentes de datos en uso incluyen procesos de compresión, cifrado y deduplicación, entendiéndose por esto último a una forma específica de compresión donde los datos superfluos son eliminados.

CAPÍTULO 6.- Sistemas tradicionales vs sistemas virtualizados

6.1.- Sistemas tradicionales de almacenamiento

Los sistemas de almacenamiento tradicionales están basados en una red SAN compuesta por:

- Host: Servidor con un determinado sistema operativo en el que están instalados HBAs.
- HBA: Tarjeta interna que permite la conexión a través de fibra. Cada puerto de cada HBA se identifica a través de su WWN (World Wide Name)
- Cabina de almacenamiento: Array de discos que proporcionan los volúmenes o LUNs a los host de manera redundante.
- Fabric: Interconexión entre host y cabina de almacenamiento, generalmente a través de fibra, que proporciona seguridad y redundancia.

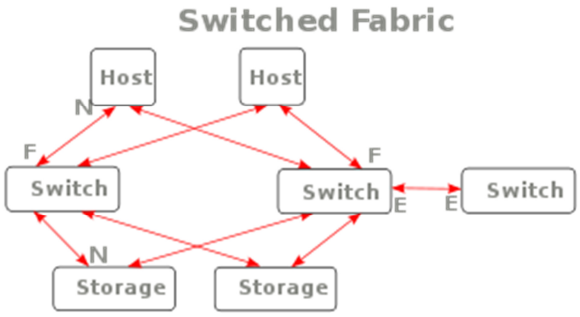


6.1.1.- Fabric

Definimos Fabric como el conjunto de switches físicos que comparten una misma asociación lógica.

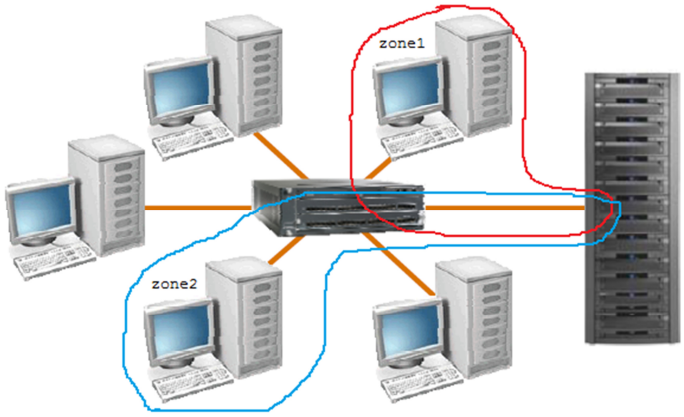
Un Fabric puede estar compuesto por uno o más switches interconectados entre sí a través de fibra mediante una conexión especial denominada ISL (Inter switch link). Mediante este link los diferentes switches que forman parte del fabric comparte toda la información entre ellos.

Por regla general se suele configurar al menos 2 fabrics compuestos cada uno de ellos por, como mínimo, dos switches. Conseguimos de esta manera tener la máxima redundancia a nivel de electrónica de red.



En una red SAN, a diferencia de una LAN donde todos los equipos pueden hablar con todos (con matices), la idea es que un servidor únicamente pueda comunicarse por Fibre Channel con su almacenamiento de manera que no haya forma de que un servidor perciba la existencia de otro. Para conseguir esto se definió el zoning.

Dentro de cada switch, de cada Fabric, existe activo un zoneset, que no es otra cosa que una lista donde se indica que servidores pueden hablar con qué equipos de almacenamiento (o con otros servidores), de manera que un switch solo da visibilidad entre sí a los equipos que comparten una zona.



Este zoneset activo (o efectivo) debe ser el mismo en todos los switchs del fabric, para mantener así la consistencia del entorno. Al conjunto de zonas, zoneset, alias y demás elementos involucrados en esto, se conoce como zoning.

Con el zoning aparecen unos nuevos términos:

- Zona: elemento que indica que equipos pueden comunicarse entre sí. Una zona puede contener, un WWN, un alias o un puerto Físico. Una zona puede tener 2 o más miembros.
- Alias: a la hora de manipular el zoning, si lo hacemos por WWN o puerto podemos confundirnos facilmente, por eso se definen los un alias, que es un nombre asociado a un WWN o puerto. De esta forma es mucho más sencillo de identificar un elemento.
- Zoneset: un zoneset es un conjunto de zonas. En un switch pueden existir varios zonesets definidos, pero únicamente puede haber uno activo.
- zoneset activo (o efectivo): es el zoneset que está activo y es el mismo para todos los switchs del fabric.

6.1.2.- Cabina de almacenamiento

Es la que proporciona las LUNs configuradas previamente. Para proporcionar redundancia, las cabinas de almacenamiento de cualquiera de los fabricantes están compuestas por dos controladoras, encargadas de gestionar las IOs, que contienen diferentes puertos para exportar los volúmenes.

En la imagen superior se pueden ver cómo de forma simétrica se tiene una controladora en el lado izquierdo (color verde) y otra controladora en el lado derecho (color rojo).

Estas procesadoras tienen la misma configuración interna a nivel de hardware y software, así como el mismo número y tipos de puertos.

Cada una de esas controladoras tiene acceso por igual a las bandejas de disco a través de un bus específico. Los disco se agrupan en bandejas de 15 o 25 discos dependiendo del tamaño de estos y cada bandeja puede estar a su vez conectada a la siguiente, de manera que se forma lo que se conoce como array de discos.

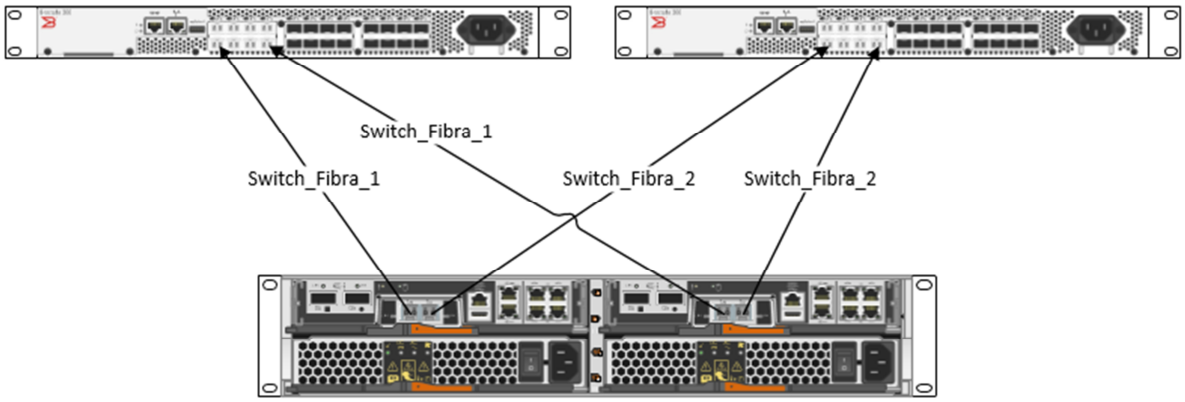
La imagen final que obtenemos es parecida a esta:

Dentro de cada bandeja de discos se pueden encontrar 3 tipos de discos:

- SAS: discos de alto rendimiento
- NL-SAS: discos de bajo rendimiento destinados, sobre todo, a archivado.
- EFD: discos de estado sólido de muy alto rendimiento.

En una misma bandeja se pueden mezclar discos de diferentes tecnologías y, tal y como se explicó en capítulos anteriores, a partir de ellos se configuran los Raid Groups/Pools y las LUNs.

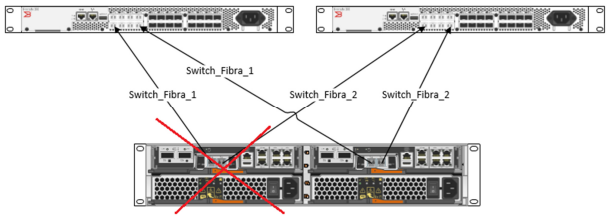
Un ejemplo aplicado de conexión redundada entre el sistema de almacenamiento y los switches es la que se muestra a continuación:



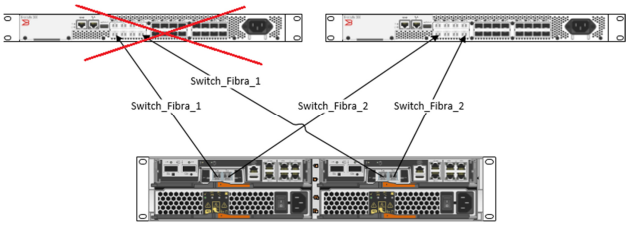
Se tienen dos fabrics compuestos por un switch de fibra cada uno. A su vez, la cabina de almacenamiento proporciona, también por redundancia, dos controladoras.

Con esa configuración estaríamos protegidos ante los posibles siguientes fallos:

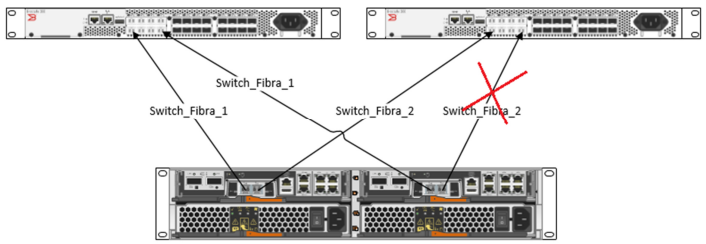
1.- Controladora de la cabina de almacenamiento:



2.- Cualquiera de los switches de cada Fabric



3.- Cualquier fallo en cualquiera de las cuatro fibras que interconectan los sistemas:



En el almacenamiento tradicional las réplicas se establecen entre dos sites remotos separados entre sí por decenas de kilómetros de forma activo/pasivo.

En este modelo solamente en site A es el principal, en el reside la producción del negocio y es replicado de manera síncrona o asíncrona contra el site B.

El volumen el site A se encuentre en estado Read/Write, por lo que cualquier host podría estar accediendo a él escribiendo o leyendo.

El volumen del site B se encuentra en estado NA, not available. No puede ser presentado a ningún host mientras se encuentre en este estado.

La recuperación ante desastres en estos casos requiere de intervención manual para cambiar el estado del volumen del site B a RW y permitir, de esta manera, acceder a los recursos.

Si decidimos una configuración síncrona obtendremos un RPO=0 y un RTO dependiente del tiempo de intervención manual para mover los recursos al site B.

6.2.- Sistemas de almacenamiento virtualizado

Tras la explicación del funcionamiento de los sistemas tradicionales de almacenamiento vamos a abordar los nuevos sistemas que nos proponen un cambio tecnológico a la hora de gestionar el acceso simultáneo a los datos.

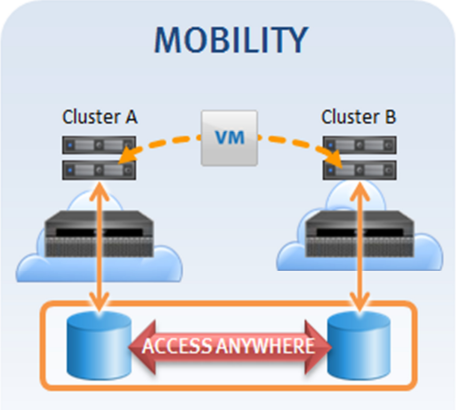
Con la virtualización del almacenamiento se busca abstraer del usuario el almacenamiento nativo para añadirle una capa virtualizada que permita trabajar con volúmenes heterogéneos como si fueran homogéneos.

La principal ventaja que nos proporciona la virtualización o federación del almacenamiento es la posibilidad de tener un sistema activo/activo real entre diferentes sites, lo que significa disponer de un RTO=0 y un RPO=0.

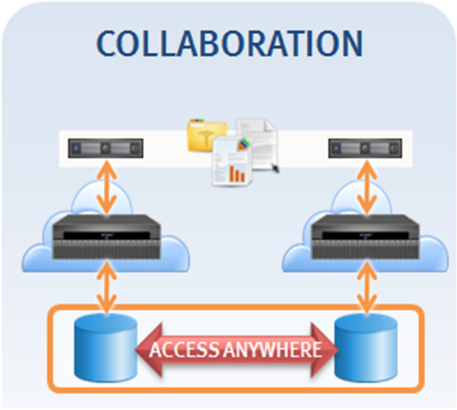
Los sistemas de almacenamiento virtualizados proporcionan las siguientes características:

1.- Disponibilidad: Con el almacenamiento virtualizado se proporciona un 100% de disponibilidad y continuidad de negocio, asegurando un RTO y RPO igual a 0. Los volúmenes permanecen continuamente sincronizados y son disponibles desde cualquiera de los dos sites aunque los separen cientos de kilómetros.

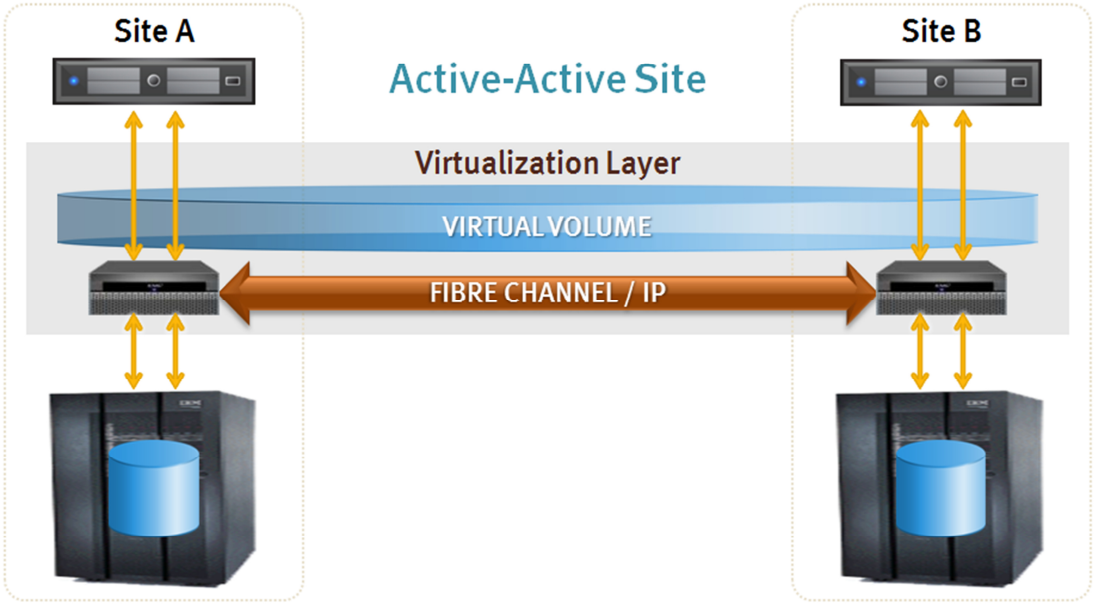
2.- Movilidad: Al contar con la capa de virtualización que abstrae los volúmenes de la cabina primaria de almacenamiento permite mover, en caliente, datos de un disco a otro sin disrupción de servicio.



3.- Colaboración: Al ser un sistema activo/activo, los datos son accesibles desde cualquiera de las dos ubicaciones de manera concurrente:



La continuidad de negocio con estos sistemas a nivel de almacenamiento está asegurada, no se requiere intervención manual para mover la producción de un site a otro.

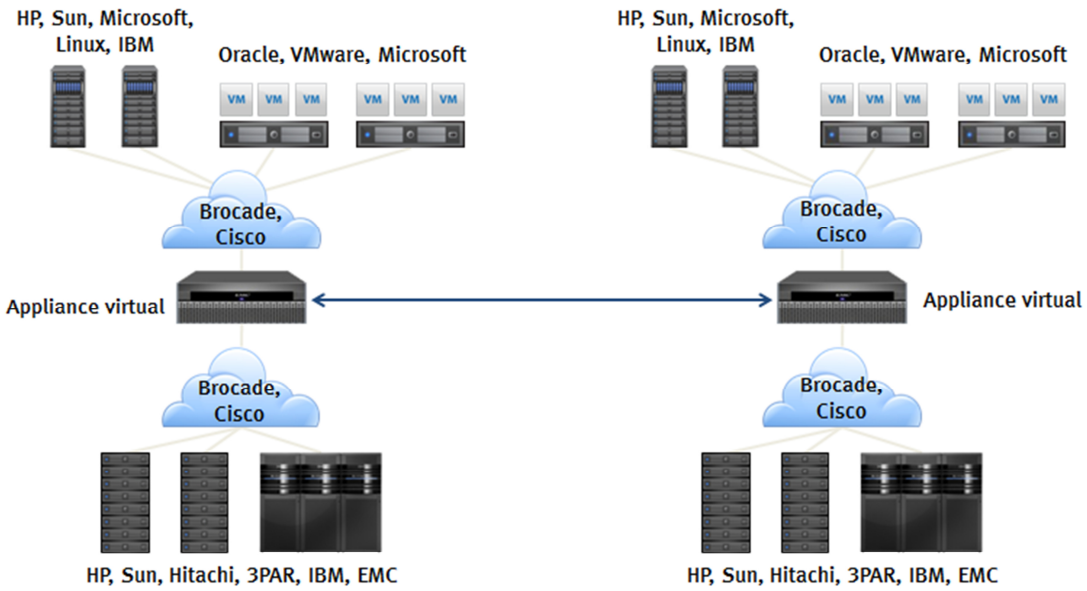


A nivel de conectividad continuaremos usando el mismo modelo que siempre a través de canales de fibra o mediante TCP/IP.

Para poder disponer de un sistema activo/activo se necesita añadir un appliance intermedio como el mostrado en la figura anterior, que tomará los volúmenes/LUNs del almacenamiento primario y los encapsulará para poder manejarlos internamente como volúmenes independientes.

Estos appliance, disponibles en varios fabricantes de almacenamiento, soportan multitud de cabinas de almacenamiento por debajo de ellos por lo que es relativamente sencillo poder montar un sistema virtualizado proporcionado por EMC, por ejemplo, con cabinas Hitachi, IBM, HP…

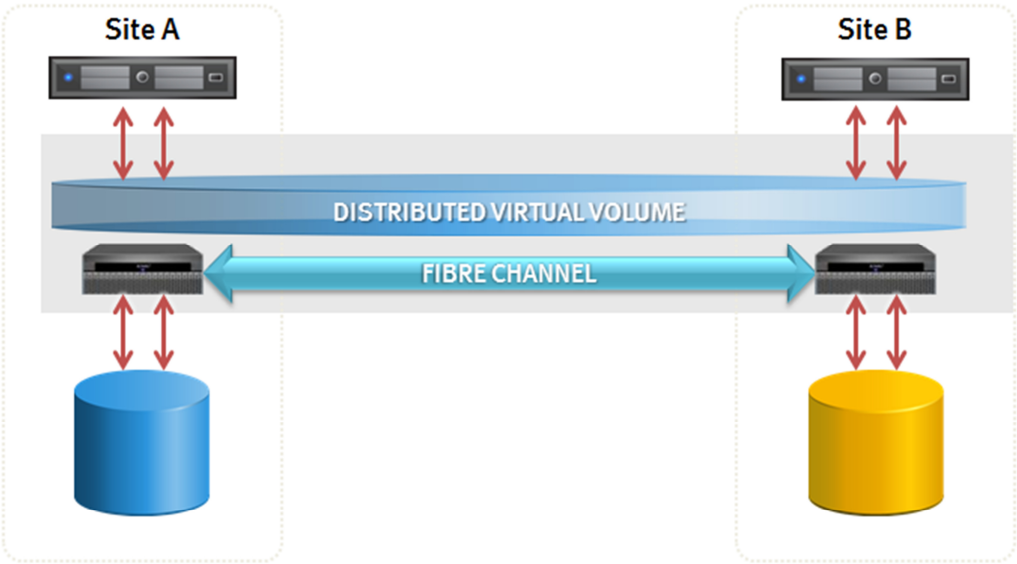
Uno de los grandes cambios de paradigma ofrecidos por estos sistemas de virtualización es, como comentábamos, la posibilidad de extender los volúmenes en un site local al site remoto manteniendo por debajo la infraestructura de almacenamiento ya existente:



Los appliance de la figura anterior se clusterizan de manera que, a nivel operativo, forman un sistema de recursos compartidos. Uno de los requerimientos más importantes para poder formar este clúster es que el RTT entre sites sea inferior a 5ms.

6.2.1.- Arquitectura de un sistema de almacenamiento virtualizado.

Cuando un disco de presenta en el site A este se virtualiza sincronizándolo con otro disco del mismo tamaño en el site B (lado derecho) formando un volumen virtual distribuido:



A nivel de gestión, cuando se crea un disco, la cabina de almacenamiento le proporciona un identificador denominado WWN (world wide name). Cuando ese disco se presenta a un host con un determinado sistema operativo, el host ve el disco con esa determinada WWN.

En el caso que nos ocupa, el disco en azul dispondrá de un WWN específico, al igual que el disco amarillo del site B.

Cuando se presentan al appliance virtual se crea un distributed virtual volumen que enmascara las WWNs de ambos discos y crea una nueva WWN. Es decir, a nivel lógico, únicamente se añade una cabecera al identificador del disco.

Por debajo, de forma transparente, es el appliance virtual el encargado de mantener bien sincronizados los volúmenes azul y amarillo del ejemplo de manera que el volumen virtualizado siempre sea consistente entre los dos sites.

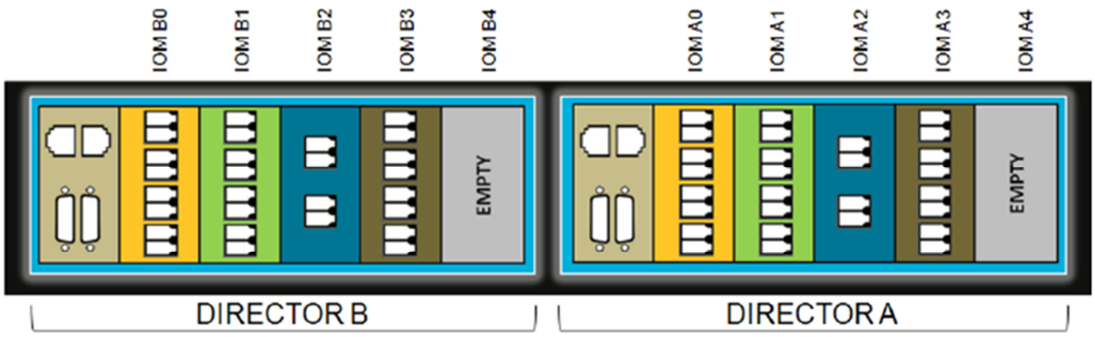
6.2.2.- Sistema de caché

Mientras que en un sistema tradicional activo/pasivo el disco de un site se encuentra en estado RW y el del site remoto en NA, en los sistemas activo/activo ambos discos son accesibles de manera concurrente desde cualquiera de los sites.

Para poder gestionar de manera correcta las reservas scsi sobre discos y los bloqueos necesarios sobre los bloques físicos se requiere de un sistema de caché muy avanzado.

Esto, junto a la combinación de un algoritmo específico permite el acceso concurrente.

Cada engine dispone de dos directores que proporcionan la alta disponibilidad que se espera de un sistema de este tipo, y cada uno de ellos con sus correspondientes puertos:



Cada uno de los directores dispone de dos tipos de caché:

1.- Caché local: Caché física donde se guardan los datos de lectura/escritura que acelerarán la consulta de datos posteriores.

2.- Caché global: Caché física que contiene la dirección física de dónde se encuentran los datos en el resto de cachés locales. Las cachés globales están sincronizadas entre todos los directores.

Es decir, en un sistema en el que tuviéramos dos sites (Cartagena y Murcia) con dos engines nos encontraríamos ante el siguiente escenario, compuesto por:

1.- Almacenamiento en Cartagena

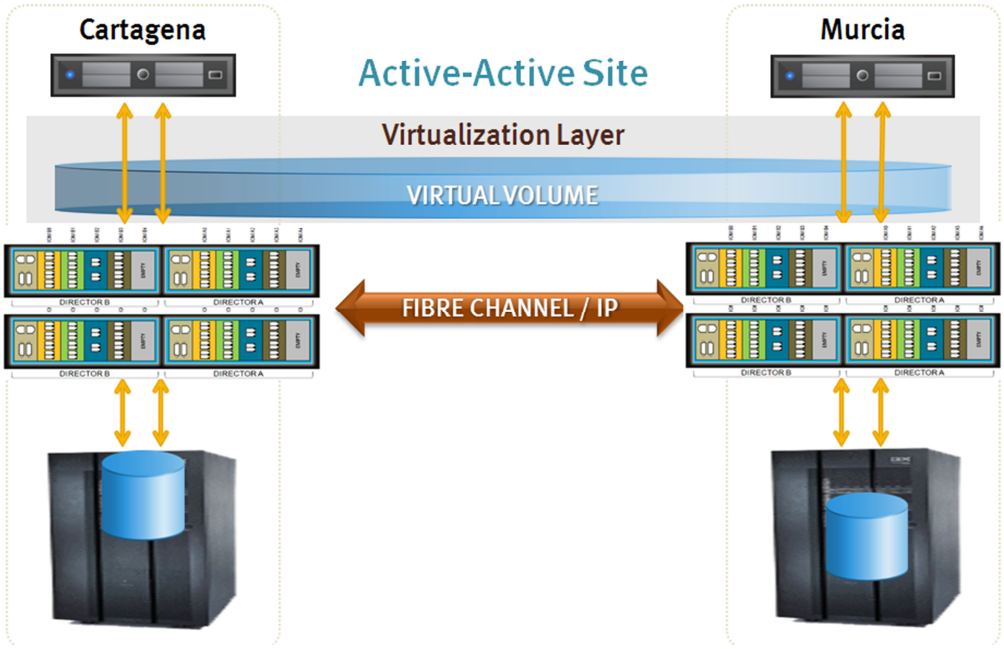
- Almacenamiento primario ya existente
- Dos engines con dos directores cada uno, sumando un total de 4 directores

2.- Almacenamiento en Murcia

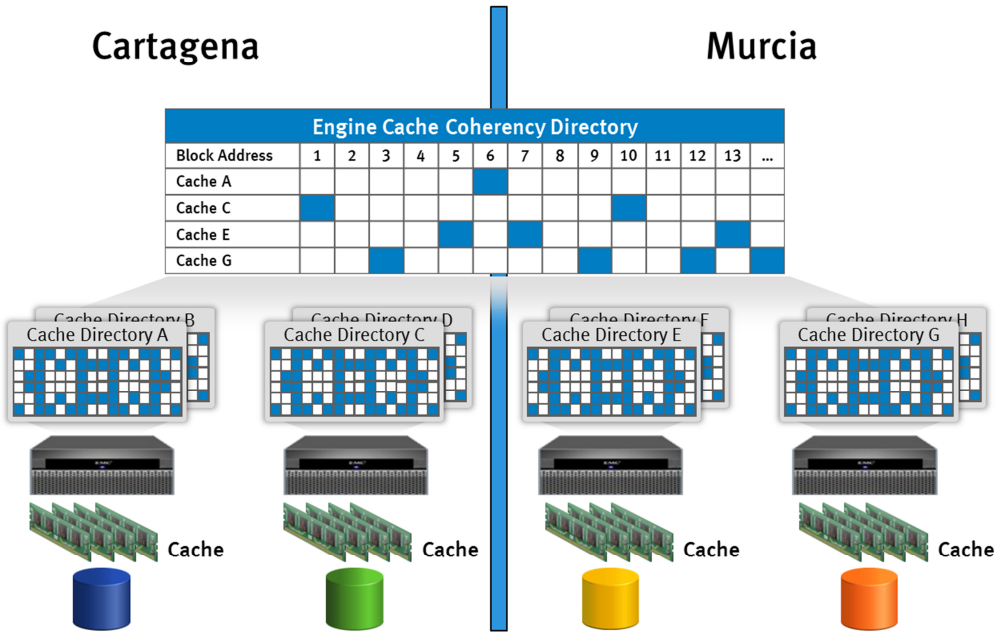
- Almacenamiento primario ya existente
- Dos engines con dos directores cada uno, sumando un total de 4 directores

La conectividad entre los dos sites se haría a través de fibra mediante switches de fibra.

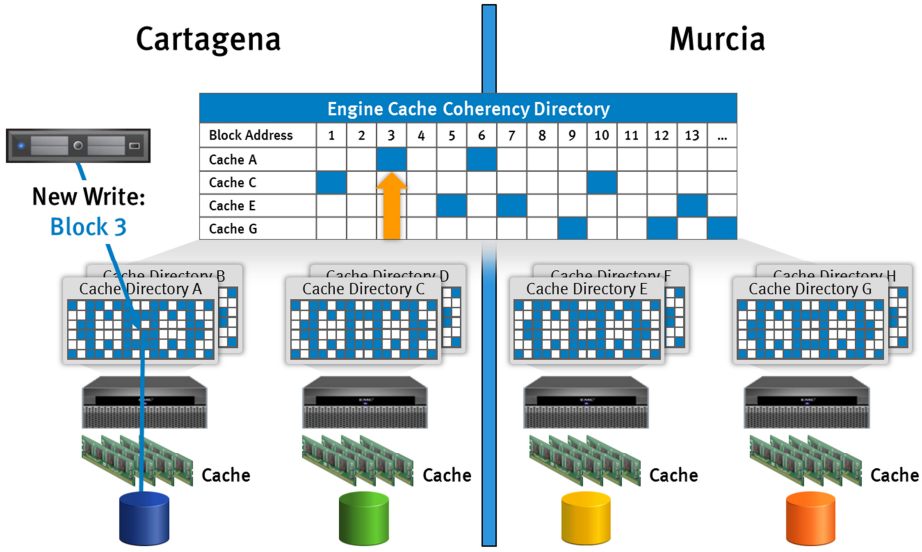
La conectividad entre los engines y el almacenamiento primario se haría a través de la SAN y sus correspondientes Fabrics.



Para permitir la escritura y lectura de manera concurrente entre los dos sites de Cartagena y Murcia dispondremos, tal y como hemos comentado, de una caché local por director y de una caché global sincronizada entre todos los directores:



Cuando una nueva escritura llega, por ejemplo del bloque 3, a través del director A del site de Cartagena, esta se queda almacenada en la caché local del correspondiente director y se apunta en la caché global dónde se encuentra físicamente ese bloque (en este ejemplo, caché A posición 3):



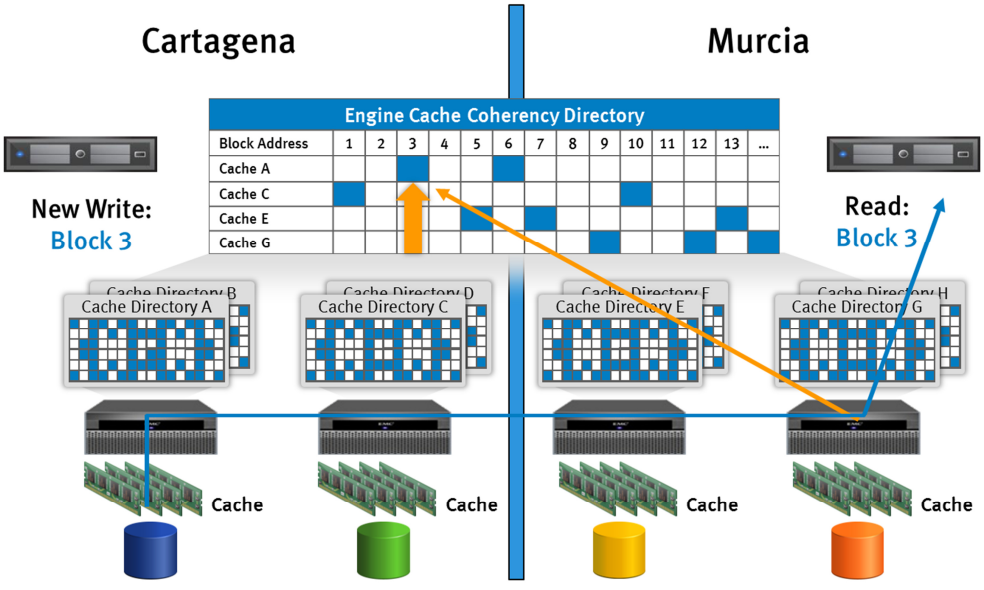
Cuando desde el site de Murcia se manda una lectura del bloque 3 (escrito anteriormente desde el host de Cartagena) el flujo que ocurre es el siguiente:

1.- Llega la escritura y entra por el director G

2.- Comprueba si tiene en su caché el bloque 3.

3.- Al no tenerlo, recurre a la caché global y pregunta por el bloque 3. De allí saca que ese dato se encuentra en el director A posición 3.

4. Se va al director que posee el dato y se lo trae, lo guarda en su caché local para futuras consultas y devuelve el dato al host que lo ha pedido.



Con esta primera aproximación al funcionamiento de la caché local y distribuida podemos hacernos una idea de cómo trabajan los sistemas virtualizados para permitir el acceso concurrente a los datos.

Si en algún momento se quiere escribir en un bloque están siendo escrito en ese mismo momento la IO se encola hasta que se libera el bloqueo.

Ya que el mantenimiento de las cachés se basa en datos estadísticos, tenemos 4 diferentes estados que nos condicionarán el mayor o menor rendimiento a la hora de escribir o leer en un sistema virtualizado.

6.2.3.- Tolerancia a fallos

Una de las mayores novedades y ventajas que proporcionan los sistemas virtualizados de almacenamiento es su fortaleza antes fallos del entorno y que solucionan gracias a su sistema Activo Activo.

Como hemos descrito anteriormente, estos sistemas proporcionan un RTO y un RPO igual a 0 en cualquiera de las circunstancias.

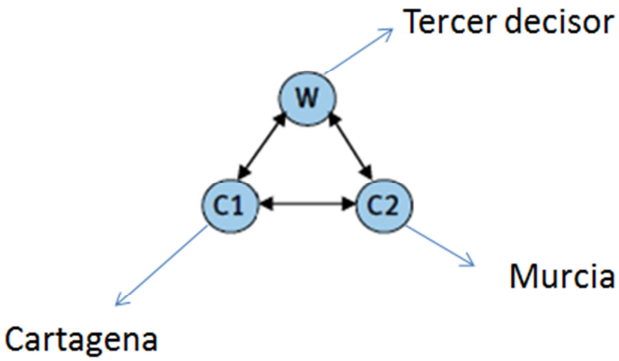
Recordemos que estos sistemas crean un único volumen virtual formado por dos componentes, dos LUNs, separadas entre sites y sincronizadas entre sí:

Cuando ocurre un fallo de site completo, Cartagena por ejemplo, el componente 1 del volumen virtualizado queda no disponible pero el host puede seguir escribiendo en ese mismo disco, en el componente 2, ubicado en Murcia.

Cuando se recupera el site de Cartagena los dos componentes se vuelven a sincronizar en el sentido Murcia -> Cartagena y el sistema vuelve a estabilizarse.

Los escenarios que pueden aplicar a catástrofes o sistemas de contingencia son múltiples y variados. Para minimizar al máximo los escenarios se puede, por ejemplo, añadir un tercer decisor, un tercer voto, que tenga una visibilidad global del sistema completo y pueda decidir en caso de duda. Este voto puede ser un sistema software, una máquina virtual, ubicada en un tercer site, que tenga conectividad ip entre todos los sistemas clusterizados.

Esquematizando los sistemas involucrados en la virtualización del almacenamiento tenemos:



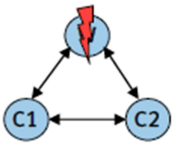
C1: Site de Cartagena

C2: Site de Murcia

W: Tercer decisor

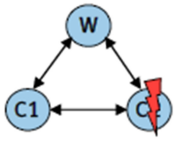
Los escenarios que podemos tener son:

1.- Fallo del tercer decisor.



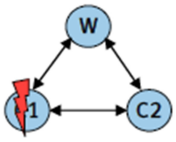
El sistema permanece online, el volumen virtual es accesible desde cualquiera de los dos sites de forma activa. El fallo del tercer decisor provoca que haya 2 votos frente a 1.

2.- Fallo del site completo de Murcia



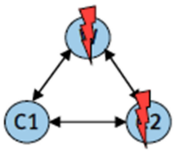
El fallo del site de Murcia provoca que el componente del volumen virtual que permanece en ese site deje de dar servicio. Los host siguen accediendo al volumen virtual escribiendo y leyendo desde el site de Cartagena. Una vez que se reestablezca el site de Murcia los dos componentes se sincronizarán y volveremos a tener un activo/activo.

3.- Fallo del site completo de Murcia



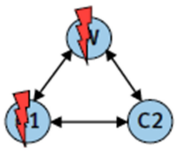
Estamos ante el mismo caso que en el punto 2 pero con el fallo de site de Cartagena en lugar del de Murcia.

4.- Fallo del site de Murcia y el tercer decisor



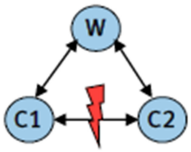
Tenemos un fallo del site completo de Murcia y del tercer decisor. La suma de votos no válidos es igual a 2, válidos igual 1. Por tanto el sistema no puede continuar dando servicio. Los host no pueden acceder al volumen virtual ya que para asegurar la consistencia de datos se para el sistema.

5.- Fallo del site de Cartagena y el tercer decisor.



Mismo caso que el punto 5 con el site de Cartagena.

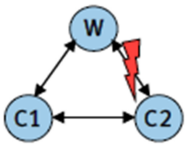
6.- Fallo de la línea de comunicaciones entre los sites de Murcia y Cartagena



Estamos ante el caso de aislamiento de sites por un corte en la electrónica del DWDM. En este escenario, cada clúster da servicio a los discos que se le haya asignado. Es decir, si a un host se le ha especificado que su ganador es Cartagena, el componente que resida en el site de Cartagena seguirá dando servicio mientras su otro componente de Murcia quedará parado.

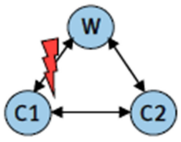
Para el host al que se haya establecido que su ganador es Murcia ocurrirá lo mismo. Una vez que se reestablezca la comunicación entre sites, los componentes se sincronizarán de nuevo.

7.- Pérdida de comunicación entre el tercer decisor y el site de Murcia



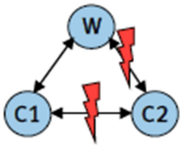
Los dos sites siguen dando servicio.

8.- Pérdida de comunicación entre el tercer decisor y el site de Cartagena



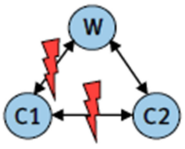
Los dos sites siguen dando servicio.

9.- Pérdida de comunicación ente el tercer decisor y el site de Murcia. Aislamiento de los dos sites



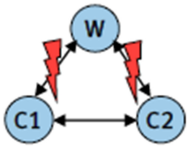
El site de Cartagena continúa dando servicio ya que tiene dos votos y el site de Murcia quedará offline.

10.- Pérdida de comunicación ente el tercer decisor y el site de Cartagena. Aislamiento de los dos sites



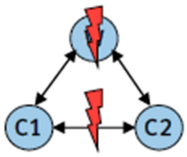
Mismo caso que el punto 9 pero con el site de Murcia.

11.- Pérdida de comunicación entre el tercer decisor y los sites de Murcia y Cartagena



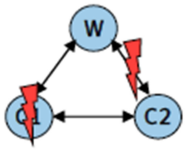
Los dos sites siguen dando servicio ya que el número de votos vivos sigue siendo mayor al de votos no disponibles.

12.- Fallo del tercer decisor y aislamiento de los sites de Murcia y Cartagena



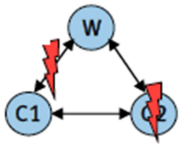
Ambos sites se paran ya que no hay votos suficientes para seguir dando servicio.

13-. Fallo del site de Cartagena y pérdida de comunicación entre el tercer voto y el site de Murcia



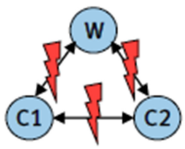
Ambos sites se paran, no hay votos suficientes para continuar dando servicio en el site de Murcia.

14.- Fallo del site de Murcia y pérdida de comunicación entre el tercer voto y el site de Cartagena



Mismo escenario que en el punto 13.

15.- Fallo completo de comunicaciones



Estamos antes el peor de los escenarios. Todas las comunicaciones se han perdido y, por tanto, no hay votos suficientes para seguir dando servicio. Toda la producción se para.

CAPÍTULO 7.- Costes de la virtualización

Aunque las mejoras derivadas de la virtualización del datacenter son elevadas los beneficios que reportan son suficientemente elevados como para estudiar profundamente su implantación.

En los siguientes puntos revisaremos los costes que se asocian, de forma genérica, a los diferentes entornos de datacenter existentes.

7.1.- Costes asociados al almacenamiento tradicional.

Para los costes asociados al almacenamiento tradicional tendremos en cuenta el coste del hardware, del software y sus correspondientes licencias, el coste de la garantía del equipamiento y el consumo eléctrico en cada uno de los sites:

Site 1:

Hardware Sub-total	**€149.843,54**
Software Sub-total	**€45.353,71**
Prepaid SW Maintenance Sub-total	**€9.948,60**
Warranty and Warranty Upgrades Sub-total	**€13.125,24**
Configuration Total	**€218.271,09**

Site 2:

Hardware Sub-total	**€149.843,54**
Software Sub-total	**€45.353,71**
Prepaid SW Maintenance Sub-total	**€9.948,60**
Warranty and Warranty Upgrades Sub-total	**€13.125,24**
Configuration Total	**€218.271,09**

Coste asociado a la electrónica de red consistente en 2 switches de fibra por site conectados entre sí formando dos fabrics:

Hardware Sub-total	**€36.181,78**
Warranty and Warranty Upgrades Sub-total	**€1.578,96**
Configuration Total	**€37.760,74**

Coste eléctrico de poner en funcionamiento todo el equipamiento asumiendo que:

- Ratio local del KW = 0.15 € / KW-hr
- Voltaje = 220v

	Power Consumption (kVA)	Heat Dissipation (Btu/hr)	Annualized Energy Cost
Almacenamiento mid-range CPD1 System Total	1,57	4.800	€ 3.698
Almacenamiento mid-range CPD2 System Total	1,57	4.800	€ 3.698
Switches System Total	0,22	800	€ 578
Site Total	3,36	10.400	€ 7.974

El coste total del almacenamiento utilizado en un datacenter en una configuración activo/pasivo replicada entre sí de manera síncrona asciende a:

Concepto	**CPD1**	**CPD2**	**Switches**
Hardware	€218.271,09	€218.271,09	€36.181,78
Software	€45.353,71	€45.353,71	
Warranty	€23073,84	€23073,84	€1.578,96
Sub-Total	€218.271,09	€218.271,09	€37.760,74
Energy cost	€3698	€3698	€578
Total	**€482276,92**		

7.2.- Costes asociados al almacenamiento virtualizado

Los costes asociados al almacenamiento virtualizado tienen dos enfoques diferentes dependiendo de si la virtualización de realiza de un entorno ya existente al que solamente se le añade la capa de virtualización o de si, por el contrario, hay que añadir también la capa de almacenamiento como tal.

7.2.1.-Virtualización de un entorno ya existente

Para este escenario se asume que ya se dispone de todas las cabinas de almacenamiento y que únicamente se va a proceder a añadir la capa de virtualización que será la encargada de abstraer el almacenamiento tradicional tal y como hemos ido explicando en los capítulos anteriores.

Site 1:

Virtualization layer	
Hardware Sub-total	€46.009,04
Software Sub-total	€95.163,54
Prepaid SW Maintenance Sub-total	€51.388,56
Warranty and Warranty Upgrades Sub-total	€0,00
Configuration Total	**€192.561,14**

Site 2:

Virtualization layer	
Hardware Sub-total	€46.009,04
Software Sub-total	€95.163,54
Prepaid SW Maintenance Sub-total	€51.388,56
Warranty and Warranty Upgrades Sub-total	€0,00
Configuration Total	**€192.561,14**

Coste eléctrico total de la solución:

	Power Consumption (kVA)	Heat Dissipation (Btu/hr)	Annualized Energy Cost
Virtualizador de almacenamiento en CPD1 System Total	0,60	1.900	€ 1.451
Virtualizador de almacenamiento en CPD2 System Total	0,60	1.900	€ 1.451
Site Total	1,20	2.800	€ 2.902

Concepto	**CPD1**	**CPD2**
Hardware	€46.009,04	€46.009,04
Software	€95.163,54	€95.163,54
Warranty	€51.388,56	€51.388,56
Sub-Total	€192561,14	€192561,14
Energy cost	€1451	€1451
Total	**€388024,28**	

7.2.2.-Virtualización de un entorno completo

Para virtualizar el datacenter completo deberemos sumar los resultados del punto 7.1 y los del punto 7.2.1, quedando de la siguiente manera:

Costes de almacenamiento base:

Concepto	**CPD1**	**CPD2**	**Switches**
Hardware	€218.271,09	€218.271,09	€36.181,78
Software	€45.353,71	€45.353,71	-
Warranty	€23073,84	€23073,84	€1.578,96
Sub-Total	€218.271,09	€218.271,09	€37.760,74
Energy cost	€3698	€3698	€578
Total	**€482276,92**		

Coste de la virtualización del almacenamiento:

Concepto	**CPD1**	**CPD2**
Hardware	€46.009,04	€46.009,04
Software	€95.163,54	€95.163,54
Warranty	€51.388,56	€51.388,56
Sub-Total	€192561,14	€192561,14
Energy cost	€1451	€1451
Total	**€388024,28**	

Costes totales:

Sub-Total 1	€482276,92
Sub-Total 2	€388024,28
Total	**€870301,2**

7.3.- Resumen de costes

Cruzando los datos obtenidos en el punto 7.1 y 7.2 con la aplicación del RTO/RPO calculada en el punto 5.2 podemos ver que:

La implantación de un sistema tradicional de almacenamiento activo/pasivo replicado de manera síncrona con un RPO=0 y un RTO dependiente del tiempo en efectuar el movimiento de datos y servidores de un datacenter a otro es de €482.276,92

La implantación de un sistema virtualizado de almacenamiento activo/activo con un RPO=0 y un RTO=0, que nos asegura una continuidad de negocio sin disrupción es de €388.024,28 en el caso de añadir únicamente la capa de virtualización y de €870.301,2 en el caso en el que se implementa toda una solución virtualizada.

Es importante añadir que todos los costes asociados al hardware, software y su correspondiente soporte técnico reflejados en las tablas anteriores no cuentan con ningún tipo de descuento. Dependiendo del tipo de cliente, el tipo de sector y otros muchos parámetros los descuentos pueden llegar hasta un 80% del valor establecido por el proveedor.

Los costes asociados al ejemplo del punto 5.2 de una empresa con una facturación anual de €36.060.726,26, con un ingreso por hora de 11.382,81 €, con un coste total promedio por empleado/hora de 15€. Los costes total por hora ascienden a:

	€/hora	Por 36 horas
Pérdida de ingresos	10.820 €	389.500 €
Costes personal	2.250 €	81.130 €
Costes intangibles	4.250 €	151.456 €
Coste total del fallo	17.300 €	622.047 €

A la vista de los resultados es tarea de las organizaciones valorar la inversión en tecnología y reputación que proteja la continuidad de su negocio.

Un fallo de 36 horas debido a comunicaciones, desastres naturales, fallo humano… ya casi duplica la inversión que debe realizar para virtualizar su datacenter y conseguir un RTO/RPO igual a 0.

CAPÍTULO 8.- Bibliografía y referencias

La bibliografía básica utilizada en la redacción de este proyecto se basa en los siguientes documentos y libros técnicos:

- Storage Technology Foundation http://www.emceducation.co.kr/file/STF.pdf
- Information Storage and Management
- Administración de Storage y Backups, Ra-Ma Editorial
- Wikipedia